IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 19, 2021, accepted October 17, 2021, date of publication October 21, 2021,
date of current version November 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122025

Urdu Sentiment Analysis via Multimodal Data Mining Based on Deep Learning Algorithms

UROOBA SEHAR[1], SUMMRINA KANWAL[2], KIA DASHTIPUR[3], USAMA MIR[4], (Senior Member, IEEE), UBAID ABBASI[5], AND FAIZA KHAN[1]

[1]Faculty of Computing, Riphah International University, Islamabad 45211, Pakistan
[2]Department of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia
[3]James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, U.K.
[4]School of Computer Science, University of Windsor, Windsor, ON N9B 3P4, Canada
[5]Department of Sciences, GPRC, Grande Prairie, AB T8V 4C4, Canada

Corresponding author: Faiza Khan (khanfaiza706@gmail.com)

ABSTRACT Every day, a massive amount of text, audio, and video data is published on websites all over the world. This valuable data can be used to gauge global trends and public perceptions. Companies are showcasing their preferred advertisements to consumers based on their online behavioral trends. Carefully analyzing this raw data to uncover useful patterns is indeed a challenging task, even more so for a resource-constrained language such as Urdu. A unique Urdu language-based multimodal dataset containing 1372 expressions has been presented in this paper as a first step to address the challenge to reveal useful patterns. Secondly, we have also presented a novel framework for multimodal sentiment analysis (MSA) that incorporates acoustic, visual, and textual responses to detect context-aware sentiments. Furthermore, we have used both decision-level and feature-level fusion methods to improve sentiment polarity prediction. The experimental results demonstrated that integration of multimodal features improves the polarity detection capability of the proposed algorithm from 84.32% (with unimodal features) to 95.35% (with multimodal features).

INDEX TERMS Multimodal sentiment analysis (MSA), Urdu sentiment analysis (URSA), convolutional neural network (CNN), long short-term memory (LSTM).

I. INTRODUCTION

The advent of social media platforms has facilitated the spreading of knowledge and opinions on a variety of global issues. People use the internet to exchange information, opinions, and feelings about products, events, services, and political issues. Social media communication platforms such as Twitter, Instagram, YouTube, and Facebook enable people to discuss a wide range of subjects, issues, and challenges and make them express themselves in a variety of ways, such as through text, images, and videos. This abundance of freely available information has resulted in the development of intelligent sentiment analysis tools to assist firms, institutions, and businesses in making more informed decisions [1]. Our proposed algorithm serves a variety of purposes, that has been described in detail in the next sections.

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

To date, the majority of research has concentrated on sentiment analysis (SA) of textual data in the English language. However, with the advent of social media, people are increasingly inclined to express their sentiments through other means as well, including images, videos, and audio in their native languages. There is, therefore, a need to apply SA to other languages as well to avoid overlooking critical information that might be presented in alternate languages and modalities.

Urdu is Pakistan's official national language and is spoken in South Asia more frequently [38]. There is a wealth of content available in Urdu on the internet, which is posted in a variety of formats, including text, audio, and visual. Many native Urdu speakers prefer to express themselves in Urdu. Our primary motivation for conducting this research is to ascertain the polarity of people's opinions which is expressed in the Urdu language on various online platforms regarding several topics.

There have been few research studies on conducting sentiment analysis of the URDU language using multimodal

data. This is due to the lack of interest on the part of Urdu language revival authorities and a scarcity of linguistic resources. Numerous studies have concentrated on text-based SA [37], in which data is gathered solely from written text. This text could be a Facebook status update or a comment, a Twitter tweet, or a film review. Using words only to regulate a person's emotion may result in ineffective consequences, as the context has a significant impact on the meaning; for example, sarcastic and other forms of mocking languages are difficult to determine.

We have explored the effects of the multimodal Sentiment Analysis (MSA) framework for Urdu language sentiment analysis, as described in [2], [3] and illustrated in Fig. 1.

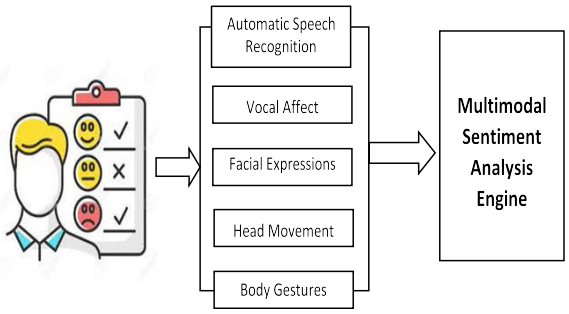

FIGURE 1. A generic multimodal sentiment analysis framework [3].

To determine the divergence of a particular video segment, the model makes use of extracted visual, audio, and textual features from the video. The video is transcribed to text in this model, and then the audio characteristics are extracted using openSMILE. Additionally, visual features are extracted. Finally, all extracted features are combined to determine the segment's complete polarity. Furthermore, we have included a dataset for Urdu MSA that has been used to conduct the experimentation for this study and can be used by other researchers for future research.

This article's overall contribution can be summarized as follows:

- A novel application of the SA framework to extract the unimodal as well as the multimodal features through a convolutional neural network (CNN) and long short-term memory (LSTM) for the Urdu language.
- Development of a multimodal Urdu language dataset collected from YouTube, comprising people's opinions in the Urdu language. The dataset is analyzed and annotated for SA implementation.
- Employment of the proposed multimodal framework for determining sentiment polarity from Urdu videos available online, and a comparison with the test-based SA approach.
- Presentation of some important experimental results to ensure a contextually integrated set of the visual, audio, verbal, and text data utilized for SA.

The rest of this paper is organized as follows: Section II presents the literature review. Section III describes the methodology and the proposed framework for Urdu MSA. Section IV discusses the dataset that we have collected for our experimentation, Section V describes the experimental results. Finally, Section VI concludes this paper and also suggests topics for future work in the area.

II. LITERATURE REVIEW

This section addresses recent researches in the automated SA domain that have been conducted by a variety of researchers. The research reveals that considerable work has already been done in the domain of SA using a variety of methods, including context-aware multimodal solutions for a variety of languages. Caon *et al.* [6] presented a framework for the sharing of multimodal emotions for both humans-to-computer and computer-to-human interactions in different social media networks, respectively. Both inputs and outputs for the multimodal context of their framework were divided into a smart environment for providing an effective way of communication and a more natural experience for the interaction. For improving the quality of the feedback, the authors have used context information. They have also implemented an evaluation scenario and conducted an observational study during interactions with the participants in the experiment. Dashtipour *et al.* [1] have presented a multimodal framework specifically for the SA of the Persian language that concurrently explores audio, visual, and textual features for a more accurate determination of the expressed multimodal sentiment. The experimental results of their research show that if the multimodal features are contextually integrated, then it can result in better performance (up to an accuracy level of 91.39%) as compared to the features that are extracted by a unimodal (having an accuracy value of up to 89.24%). Chauhan *et al.* [7] have devised a multimodal emotions analysis model based on the recurrent neural network (RNN). This model also learns the interaction within different participating models by using an auto-encoder algorithm-based mechanism. Mukhtar *et al.* [8] presented an Urdu language-oriented SA model that tests three different algorithms, namely, the decision tree (DT), the k-nearest neighbor (KNN), and the support vector machine (SVM). The outputs of all these algorithms have been compared and improvements in their results have been achieved using various processes such as the stopping word removal and the feature extraction. Mukhtar *et al.* [9] used various SA techniques and carried out SA of Urdu blogs from several fields, such as the lexicon-based approach and the supervised ML approach. In the Lexicon-based approach, they have used the Urdu sentiment lexicon as well as an efficient analyzer for Urdu sentiments. The best accuracy achieved was 67.02% using KNN as the most appropriate classifier. Mehmood *et al.* [10] proposed a novel approach that they call encoding based on transliteration for Roman Hindi or Urdu text normalization (as TERUN). The TERUN model consists of three inter-associated modules, namely, an encoder based on transliteration, a module for filtration, and a ranker for the hash codes. The encoder can generate all the hash codes that are possible for one word in Roman Hindi or Urdu. The second

module then filters out the irrelevant and unnecessary codes from the generated hash codes, and the last module finally ranks the hash-codes that are filtered out and finalized based on their applicability. The extracted results show a better efficiency than the phonetic algorithms that were well-known and widely used. Ghulam *et al.* [11] presented a long-short time memory model (LSTM) for the analysis of sentiments from Roman Urdu text. Their presented framework achieves an accuracy level of 0.95 and an F1 score of 0.94, respectively. Mahmood *et al.* [12] proposed a deep learning model for extracting the emotions and behavior of people, as expressed in the Roman Urdu language. For their experiments, they have used a dataset consisting of 10,021 sentences from 566 online threads belonging to different genres including Sports, Software, Food & Recipes, Drama, and Politics. The study shows that the recurrent CNN model provides an accuracy of 0.652, outperforming the accuracy level of the baseline models for binary classification. The proposed model also achieves an accuracy of 0.572 for tertiary classification. Furthermore, Q. Rajput [13] presented a framework for semantic annotation that can annotate documents written in the Urdu language. The proposed model used field-oriented ideology and context keywords rather than natural language processing-based techniques. The dataset was derived from the online ads that were published in digital Urdu newspapers. Rosas *et al.* [14] introduce a model that combined visual, audio, and text features to identify sentiments from online videos. Their results proved that the joint usage of text, audio, and visual features can indeed enhance the accuracy levels which is a huge advantage over the single modality-based models. The authors have also tested the portability of their proposed multimodal technique and run assessments on another dataset containing English language videos. Poria *et al.* [15] presented a comparative study focusing mainly on using audio, visual, and text features for multimodal SA along with an extensive number of studies for multiple types of fusion techniques. A detailed analysis of the improvement in the performance using multimodal analysis in comparison to the techniques with single modality analysis was also given. Nawaz *et al.* [16] developed a framework for the automated generation of extractive summaries. In their developed framework, approaches based on local and global weights were used for the Urdu language. For sentence weighting, as a baseline, the vector space model (VSM) was adapted. The experiments show the approaches based on LW provide better results for an extractive summary generation where the F-scores for the sentence weighting and weighted term-frequency methods were about 80% and 76%, respectively. Gan *et al.* [17] proposed an architecture for a scalable multi-channel dilated joint CNN and a bidirectional long short-term memory (BLSTM) model with an attention technique for analyzing the sentiment capability of Chinese texts. Farha *et al.* [18] tested the use of transformer-based language models for the SA of Arabic. They have shown a performance improvement. The best model has achieved F-scores of 0.69, 0.76, and 0.92 on the datasets including SemEval, ASTD, and ArSAS. Smetanin *et al.* [19] fine-tuned the multilingual bidirectional encoder representations from Transformers (BERT) known at the RuBERT and obtained two versions of the multilingual universal sentence encoder. They reported promising results on seven different datasets of sentiments in the Russian language. Kumar *et al*. [20] proposed a hybrid, multimodal deep learning model to predict sentiments. The accuracy of the proposed model was approximately 91%. The SA technique for social media analysis was suggested by Alaoui *et al*. [21], who extracted a positive outcome from the users' real-time opinions. Bhuiyan *et al.* [22] presented an SA model based on natural language processing (NLP) for the user's feedback. They demonstrated the success of their approach by conducting a data-driven experiment examining the accuracy of identifying relevant, popular, and high-quality videos through a study of users' comments. Krishna *et al.* [23] proposed some machine-learning techniques based on the SA of YouTube comments related to popular topics. The results showed how the trends in user's sentiments are closely related to the real-world events that are being associated with the respective keywords. Syed *et al.* [24] presented an approach for SA based on the identification and the extraction of the SentiUnits from a given text, using shallow parsing. The proposed model achieved an accuracy level of around 72% on one product and 78% on another, respectively. Li *et al.* [25] proposed a cognitive brain limbic system (HALCB) based on the hierarchal attention-BiLSTM model. While compared to several baseline approaches, the authors achieved a 15% increase in accuracy when using tri-modalities.

Arao *et al.* [26] proposed a method for recognizing emotions and sentiments that integrated hyperbolic space in neural network models. They added a hyperbolic output layer to existing state-of-the-art models and found that it has the potential to improve the modal's prediction accuracy. Vashishtha *et al.* [27] developed a supervised fuzzy rule-based system for multimodal sentiment classification. Their proposed technique achieved a level of accuracy of approximately 82.5 percent. Zhang *et al.* [28] proposed a quantum-based and LSTM-based modal for the MSA. They conducted experiments on two datasets for their research: MELD and IEMOCAP. Li *et al.* [29] introduced a method for MSA that utilized a multi-perspective fusion network. They researched CMU-MOSI, MOSEI, and YouTube public datasets and found that their method improved accuracy by 2.9 percent. Agarwal *et al.* [30] proposed a modal for MSA based on deep learning. They conducted research using a variety of RNN variants, including GRNN, LRNN, GLRNN, and UGRNN. Their proposed method achieved an accuracy of 78.05 percent when used with a multimodal dataset. Yao *et al.* [31] developed a technique for classifying multimodal sentiments based on transformer modal architecture and transfer learning. Furthermore, they introduced their dataset, termed MORSE for MSA. Hussien *et al.* [32] conducted a comparison of the various MSA techniques proposed by various researchers. They concluded that while MSA has a lot of work in the English language, it nevertheless lagged

in other languages. Ullah *et al.* [33] conducted a similar comparative analysis of recently presented MSA approaches. Portes *et al.* [34] introduced the 3D Residual Network in Embedded Systems MSA technique. They conducted experiments on the MOSI dataset and obtained an F1 score of 80%. Ali *et al.* [35] proposed a sentiment classification approach based on ontology and latent dirichlet allocation. They developed their model using the web ontology language and the Java programming language. By constructing adaptive trees, Rahmani *et al.* [36] proposed an LSTM-based model. Their research demonstrates that the hierarchical clustering technique that they have presented, is a superior method for grouping users within the constructed adaptive tree.

Although numerous researches, using data from publicly available internet sources, are being undertaken in the SA field, there is, nevertheless, still a lack in the research of URDU for SA. This article, therefore, proposes a method for extracting features from the Urdu videos that are based on the deep learning paradigms of CNN and LSTM. We have introduced a unique Urdu language dataset from YouTube. Along with this unimodal and multimodal technique, both early and late fusions have been used together to discover the sentiment polarity of the Urdu language for the first time.

III. METHODOLOGY

In this section, the methodology of our proposed framework for multimodal URSA has been discussed. In our model, for URSA, the first step is the contextual extraction of the textual, audio, and visual features from the dataset using different extraction approaches. Next, the effective extracted features have been passed to the model for identification of inclusive polarity of input video dataset for the final prediction. The proposed multimodal URSA framework is Fig. 2.

Our proposed modal uses the characteristics of all three modalities as sample input for the experiment, namely: text, audio, and video. These were first contextualized using a variety of tools and techniques such as BLSTM, OPENSMILE, and 3D-CNN, which will be discussed in the next sections. The input dataset was divided into three parts: 60% for training the model, 30% for testing, and 10% for validation. Then, using early and decision level fusion, we obtained our output, which was the predicted positive or negative polarity of the target dataset for SA, as illustrated in Fig. 2 above.

A. FEATURE EXTRACTION

1) TEXT DATA

As stated before, we have used a deep learning model to extract the features for enabling prediction from the text input. The contextual extraction of features from the textual input data was performed using a layered BLSTM model. Each expression was a mix of pre-trained 300-dimensional fastText word embeddings. Next, each expression was condensed into a 30-word window. The converted parameters were then put into a BLSTM model with experimentally determined parameters.

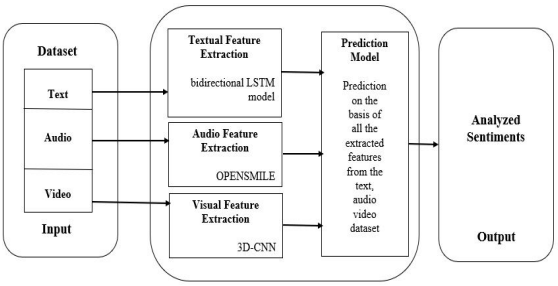

FIGURE 2. Workflow of the proposed Urdu sentiment analysis (URSA) framework.

The simplest implementation of BLSTM was proposed, specifically the one which had a two stacked BLSTM containing 128 and 64 cells, a dropout of the probability of 0.2, and 155 dense layers having two neurons and a softmax activation. The output of the last BLSTM was concatenated and passed to a fully connected layer having 128 neurons (ReLU activation) and 2 neurons (Softmax activation), respectively. With time, this complete network learned the expressions of a statement passed to it as an input. A deep learning model was used for the classification of the unclassified Urdu statements, focusing on the sentences that were not able to be categorized using rules based on the dependency. Each Urdu statement from the dataset was then transformed into a 300-dimensional vector using the fastText tool and the concatenation of word embedding was then passed to the deep learning classifiers for classification. For comparison of the results obtained from the deep learning classifier, the statements were also converted into a so-called 'bag-of-words' and passed to logistic regression (LR) and an SVM. Moreover, the transcribed videos were also converted into 300 dimensions' word embedding, which was then passed to the CNN and the LSTM for classification.

The dataset used for this experiment was divided into three parts: 50% for training the model, 25% for testing, and 25% for validation. All four models, LSTM, CNN, LR, and SVM, were trained using 50% of the dataset, then verified, and lastly evaluated using 25% of the dataset's test set. When the experimental results of various classifiers were compared, it was found out that the BLSTM combined with dependency-based rules achieved a higher level of accuracy than the other classification approaches. Thus, this approach was therefore chosen for textual feature extraction. Detailed analysis and results for these experiments were not included in this paper due to space and time constraints. The primary objective of our proposed work was to identify more resilient techniques for feature extraction from Urdu multimodal data and to apply fusion to predict opinion polarity. For this reason, we did not include any details related to the comparison of different textual/audio/video extraction techniques available. The architecture of BLSTM for textual feature extraction is depicted in Fig. 3.

2) AUDIO DATA

Several recent studies have indicated that openSMILE is quite an effective software to extract the audio features and that

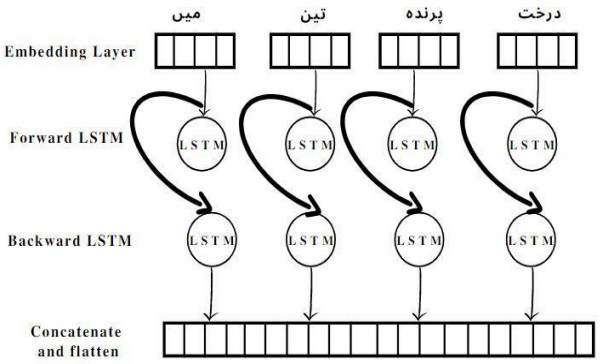

FIGURE 3. The BLSTM architecture for the extraction of textual features.

it produces some very good results. It is capable of automatically extracting the low-level descriptors from the audio recordings, even including a beat histogram, Mel frequency cepstral coefficients, a spectral centroid, a spectral flux, a beat histogram, and a beat sum. openSMILE was used in our experimentation to extract the features such as the low-level descriptors and statistical information associated with them. Moreover, some additional features such as the amplitude, arithmetic, and quadratic means, standard deviation, flatness, skewness, kurtosis, and quartiles were also extracted using openSMILE. The total number of audio features in a single sentence was 6373. These features were extracted at a rate of 40 samples per second. The normalization of speakers was accomplished by using z-standardization.

In addition to the above, the extracted features from the audio dataset were then used to analyze the sentiments employing autistic cues. The retrieved features include the following audio sub-features:

- Prosody includes loudness, intensity, and pitch that elaborates the speech signal as far as amplitude and frequency are concerned.
- The energy depicts the human loudness perception.
- Voice probabilities reveal unvoiced and voiced energies in audio.
- Spectral features use nonlinear frequency for audio stimulation.
- Cepstral features focus on the differences in the spectrum features that are measured using frequencies.

A multilayer perceptron framework (Fig. 4) was used to predict opinion polarity based on audio cues extracted from the dataset using openSMILE. Detail of the MLP architecture can be found in Table 1 below.

3) VIDEO DATA

Expressions are critical in identifying the emotions being communicated. More precisely, these are facial expressions, in conjunction with visual cues, that assist the effective identification of sentiments. Thus, visual characteristics play a critical part in multimodal SA. In our study, we employed a 'facial action coding system' to describe the facial expressions. Typically, facial expressions are classified into many

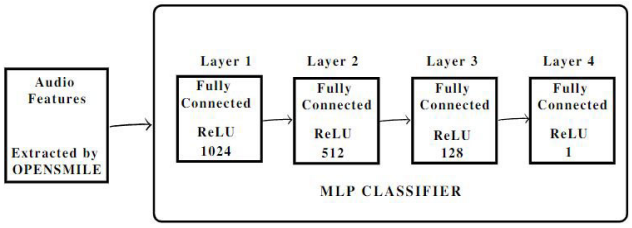

FIGURE 4. Extraction and classification of audio cues features using MLP.

TABLE 1. The MLP architecture for audio features extraction.

Layer	Type	Neuron
1	Relu	1024
2	ReLU	512
3	ReLU	128
4	ReLU	1

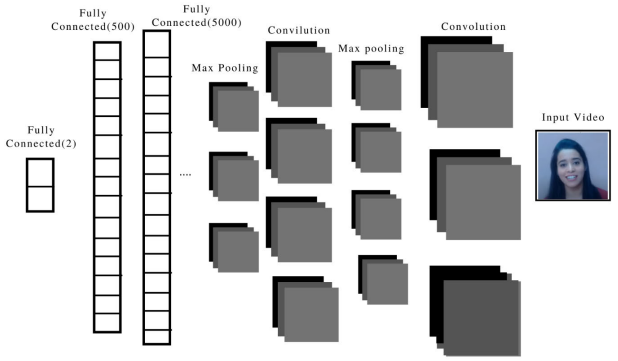

FIGURE 5. Visual features extraction from video dataset using 9-layered 3D CNN framework.

active components. However, we have extracted the visual features using a 3D CNN (three-dimensional convolutional neural network) in our proposed approach. This 3D-CNN was initiated by contextually exploring both the spatial and temporal patterns to precisely determine the Spatio-temporal relationship between a subjective and an objective expression. In our experiments, we have obtained the best results by employing a 3D CNN design with nine layers, as illustrated in Fig. 5 below. The retrieved features include not only the estimated smile and the head pose but also the facial motion units. The architecture is shown in detail in Table 2.

B. EARLY FUSION

As illustrated in Fig. 6, we began the process by extracting textual, audio, and video features using BLSTM, openSMILE, and 3D-CNN, respectively. Following this, we combined the extracted features from each input channel into a single medium before passing them to the classifiers. At this stage, low-level features from each modality were integrated, which typically resulted in increased accuracy. Early fusion captures the true spirit of this multimodal

TABLE 2. The 3d-CNN architecture for visual feature extraction.

Layer	Type		Neuron
1	3D CNN	16	2 x 2 x 2
2	3D CNN	32	2 x 2 x 2
3	3D Max Pooling	64	1 x 2 x 2
4	3D CNN		2 x 2 x 2
5	3D Max Pooling	64	2 x 2 x 2
6	3D CNN		2 x 2 x 2
7	3D Max Pooling		
8	Fully Connected	5000	1 x 2 x 2
9	Fully Connected	500	
10	Fully Connected	2	

data, improves the framework's performance, and produces superior results by combining all the features extracted and using different extractors, resulting in a single representation. In simple terms, when numerous tri-modalities are classified at various levels, the overall prediction accuracy of the modal is improved. Thus, to improve accuracy, the unimodal features were retrieved first and then fed independently to the classifiers for feature-level classification. Finally, by connecting these classifiers, we could bring all of our efforts together.

C. LATE FUSION

According to some experts, combining the classification results of individual modalities at the decision level can help to improve classification results. In late fusion, the data of each modality data is fed to the classifier for prediction, rather than merging it all at once. Finally, these predictions are combined to form a single decision vector. The individual strength of each uni-modality is the main focus of late/decision level fusions. Due to the absence of the representation problem, late fusion is easier to perform than early fusion.

For each modality, as shown in Fig. 7 below, the extracted features were classified independently. The final decision was made by combining features that have been pre-trained and were of a high level. For the predicted output, the classification results of individual modalities were combined. For the concatenation of classifiers, this late-fusion method has the advantage of requiring no "unsampling." Our multimodal prediction results were good as a result of the integration of different classifiers at different levels.

As a result of both early and late fusions, we obtained promising results. This concatenation of both levels of fusion gives us an early and late fusion advantage, which improved our modal's accuracy and efficiency.

IV. DATASET

Details of the dataset that we have used to conduct this research are given and discussed in this section.

A. OVERVIEW OF URDU LANGUAGE

Urdu is the national language of Pakistan, but it is also widely spoken in Bangladesh and India. It is a synthesis of the Persian and Arabic alphabets. Urdu alphabets contain between 39 and 40 letters. In the late 1980s,

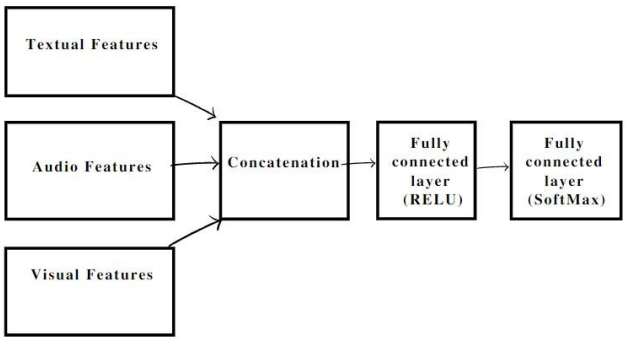

FIGURE 6. Early fusion framework of the multimodal data.

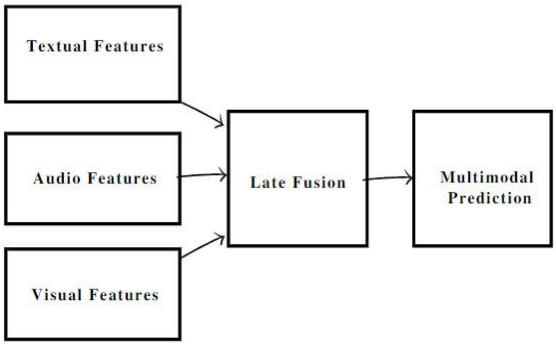

FIGURE 7. Late fusion (decision level fusion) framework for multimodal classification.

Pakistan's national daily Jung pioneered the use of computer-assisted Urdu composition (Nastaliq). Since then, more digital material in Urdu has been created. A sizable collection of free Urdu texts, graphics, and videos is now available online [4]. However, due to the scarcity of digital resources for Urdu such as sentiment lexicons, machine-readable corpus, the work of Urdu language analysis remains extremely difficult. More specific challenges to be found in the Urdu language namely, inconsistencies in case markers, and vocabulary, as well as irregularity in syntax. Furthermore, few resources are available in Urdu; hence, to date, little has been researched, recognized, and recorded about the Urdu language [5].

B. URDU MULTIMODAL DATASET

We accumulated 44 review/opinion videos (22 male/22 female Urdu speakers) from YouTube. The search for videos was based on the following keywords "book reviews in Urdu by male/man/men/boy(s)," "book reviews in Urdu by female/woman/women/girl(s)," "movie reviews in Urdu," and "cosmetic reviews in Urdu by female/woman/women/girl(s)" as well as on generic statement such as "reviews in Urdu." Furthermore, we also filtered and selected videos that satisfied the following criteria:

- The speaker should face the camera and speak clearly.
- Face visibility should be clear.
- There must be no background voices or noise.
- There is one speaker with a clear background.

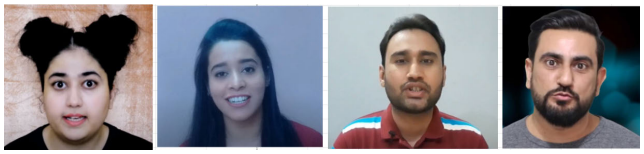

FIGURE 8. Snapshots of a few speakers from the Urdu videos are included in the experimentation dataset.

TABLE 3. Urdu multimodal dataset statistics.

	Male	Female
Number of negative utterances	250	257
Number of positive utterances	263	284
Number of objective utterances	157	161
Total number of utterances	670	702
Total number of utterances	1372	

- The video parts that had additional scenes, such as photo cover books or movie trailers, were ignored.

The final set included speakers within the 20–40-year age range and with videos that had an average length of 3 to 8 minutes. A sample snapshot of speakers from the selected Urdu videos is shown below in Fig. 8.

C. VIDEO SEGMENTATION AND TRANSCRIPTION

We excluded the part of the videos where the speakers used any English sentences. We manually selected expression of opinion utterances from each video, ensuring the start and the end of the utterance were being recorded as well. The segments were created based on the pause signs/symbols of the Urdu language such as ہیں, ہے, etc., or due to a pause by the speaker. Furthermore, the segment was then transcribed by an expert Urdu native speaker. The final transcription set contained a total of 1372 utterances (670 male/702 female). The transcription was then appraised by two Urdu language native speakers. The distribution of the video utterances with their various types in the videos is shown below in Table. 3.

D. SEGMENT ANNOTATION

Urdu language experts and native speakers annotated the utterances for their polarity as negative (−1), positive (+1), or neutral sentences (0). Agreement between all annotators was 94%, with any further disagreement being resolved through discussion. The polarity was assigned keeping in view the visual, acoustic, and textual features. Gestures such as a smile, a frown, a head nod, or a headshake were annotated manually to study the relationship between words and gestures. The gestures and utterances were manually recorded together for their polarity. On average, 10 to 20 utterances were extracted per video, with each utterance having its corresponding video and audio segmentation, respectively. In Table. 4, some examples of the utterance sentences are shown.

Three experts received the dataset and verified the correctness of the utterance categories to be subjective or objective as well as verifying their polarity. The agreement for gesture categorization reached 94%. This was carried out by marking the utterances used together with these particular gestures. Expert coders manually annotated each utterance with the gesture information. The average agreement of the gestures was 92.23%.

Due to this and other linguistic characteristics of Urdu as well as a highly opinionated web environment, we decided to collect Urdu language datasets and apply the SA framework to predict polarity using those datasets.

TABLE 4. Urdu utterance samples from the dataset.

Example sentences in Urdu	Translation in English	Polarity
تو آج کے ویڈیو میں میں اسی کا ریویو آپکے ساتھ شیر کروں گی	So, in today's video I will share its review with you.	0
دوسرا اس مووی کے اندر جو سمینہ احمدہے تو انہوں نے بہت اچھا رول کیا ہے	Secondly, Samina Ahmed has played a very good role in this movie.	1
تو میں یہ نہیں بتاسکتی ہوں کےآپکےفیس سےپمپل کو دور کرے گا کے نہیں	So, I can't say whether it will remove the pimple from your face or not.	-1
اس فیس واش کی کچھ باتیں مجھے بہت اچھی لگی	I like some of the features of this facewash.	1

TABLE 5. Classification results for audio dataset sentiment analysis.

	Precision	Recall	F-Measure	Accuracy
Positive	0.80	0.79	0.80	
Negative	0.83	0.82	0.83	
Average	0.82	0.81	0.82	89.43%

V. EXPERIMENTAL RESULTS

We applied the abovementioned methodology to predict the polarity of utterance from the Urdu multimodal dataset drawn from YouTube videos. The features were extracted contextually from the textual, audio. and video modalities. The results of applying this methodology to the unimodal dataset are given in Tables 5, 6, and 8 below.

From the tabular results above, it is evident that the audio-based features are capable of distinguishing between positive and negative utterances when compared to other modalities. This can be due to the selection of words and the tone of the speaker as well as other factors that make words more distinguishable. Furthermore, a negative utterance has a finer precision and recall value than a positive utterance.

TABLE 6. Classification results for text dataset sentiment analysis.

	Precision	Recall	F-Measure	Accuracy
Positive	0.90	0.89	0.90	
Negative	0.86	0.85	0.86	
Average	0.88	0.87	0.88	84.32%

TABLE 7. Classification results for video dataset sentiment analysis.

	Precision	Recall	F-Measure	Accuracy
Positive	0.75	0.74	0.75	
Negative	0.83	0.82	0.83	
Average	0.79	0.78	0.79	80.62%

TABLE 8. Classification of the utterance with decision level fusion.

Modality		Precision	Recall	F-measure	Accuracy
A+V	Negative	0.76	0.75	0.76	
	Positive	0.82	0.83	0.82	
	Average	0.79	0.78	0.79	80.2%
V+T	Negative	0.88	0.87	0.88	
	Positive	0.90	0.90	0.90	
	Average	0.89	0.89	0.89	92.46%
A+T	Negative	0.85	0.84	0.85	
	Positive	0.92	0.91	0.92	
	Average	0.88	0.89	0.88	89.58%
A+V+T	Negative	0.90	0.89	0.90	
	Positive	0.93	0.92	0.93	
	Average	0.92	0.91	0.92	93.32%
T-Only	Average	0.88	0.87	0.88	84.32%
A-Only	Average	0.82	0.81	0.82	89.43%
V-Only	Average	0.79	0.78	0.79	80.62%

In the case of the text-based features, the precision and the recall, as well as the F-measures, are much better, when considering the positive sentences, but the accuracy is sharper for visual features. Likewise, for the visual features, the negative utterances have more precision and recall values; compared to the negative features.

Table 8 shows that the text, vocal, and visual features provide a better prediction of the polarity of utterance when considered individually. This is because the sentiments, utterance, speaker's tone, and facial expressions are correlated when uttering a negative or positive sentence [24]. A+T (audio and textual) fusion is better than other fusions. The accuracy achieved by joining three features is much greater than when other features are involved. Furthermore, higher precision and recall values are achieved in the case of A+T for positive features than for negative ones. Similarly, even

TABLE 9. Prediction results: Early (feature-level) fusion.

Modality		Precision	Recall	F-measure	Accuracy
A+V	Negative	0.82	0.79	0.81	
	Positive	0.84	0.87	0.85	
	Average	0.79	0.82	0.80	83.33%
V+T	Negative	0.92	0.91	0.92	
	Positive	0.93	0.92	0.93	
	Average	088	0.90	0.89	94.86%
A+T	Negative	0.88	0.87	0.88	
	Positive	0.90	0.89	0.89	
	Average	0.89	0.88	0.88	89.58%
A+V+T	Negative	0.92	0.91	0.92	
	Positive	0.93	0.92	0.94	
	Average	0.92	0.91	0.93	95.85%
T-Only	Average	0.88	0.87	0.88	84.32%
A-Only	Average	0.82	0.81	0.82	89.43%

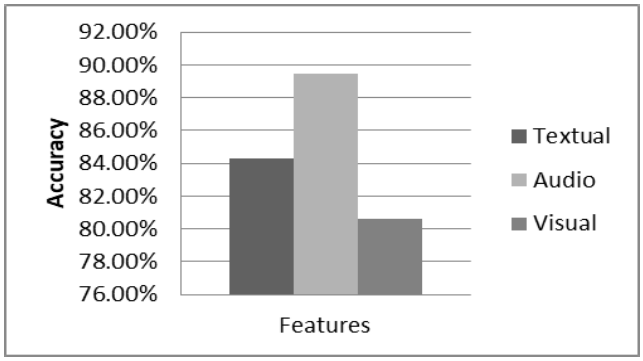

FIGURE 9. Comparison of the accuracy value achieved by unimodal analysis of Urdu dataset.

better accuracy and recall values have been reported when combining V+T (video and textual) or A+V (audio and video) for positive features.

The results generated using the feature level fusion methodology are shown in Table 9. The fusion at the feature level is more precise than the unimodal features. Fig. 9 presents the accuracy achieved by the unimodal SA for text, audio, and video modalities, while Fig. 10 depicts the performance accuracy results from early and late fusion approaches. Overall, the precision and recall of positive utterances are higher for negative utterances.

As discussed in Section II, different researchers adopt a variety of methods for various languages. In Table 10, we compare a few of the existing models for Urdu MSA. Our proposed method outperforms all other Urdu SA analysis algorithms currently available. Section 2 describes these methods in greater detail.

DISCUSSION

We have created an Urdu multimodal dataset, by collecting the review/opinion videos of male and female speakers

TABLE 10. Comparison of results with other approaches.

	Paper Title	Accuracy
Our proposed MSA Framework	A Novel Multimodal Framework for Urdu Sentiment Analysis	91.23% accuracy for decision fusion and 95.35% for feature fusion
KNN	Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains [9]	62.07%
VSM, LW	Extractive Text Summarization Models for Urdu Language [16]	80% and 76%
RNN Variants	Multimodal Sentiment Analysis via RNN variants [30]	78.05%
CNN	Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network model [12]	57.2%

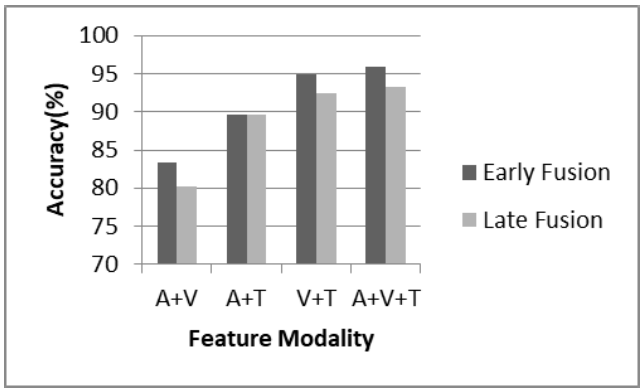

FIGURE 10. Multimodal sentiment analysis's accuracy: a comparison of early and late fusion.

from YouTube. Furthermore, we extracted the textual, audio, and video segments and derived relevant features from the videos. We then applied the proposed methodology on the unimodal as well as on the multimodal datasets and computed the prediction accuracy of the proposed framework in terms of accuracy, precision, recall, and F-measure. Experimental results showed that the proposed model improves the fusion both at the feature level and the decision level. We achieved a 91.23% accuracy in decision fusion and 95.35% in feature fusion, respectively. The combination of audio, text and visual features also showed a better degree of precision. When we look at the other result combinations, the A+T features are comparatively better compared to the related fusions where the unimodal precision performance only improves with the textual data. There were, however, still some limitations to our methodology; these were the following:

- The chosen videos were of informal speakers with significantly different expressions and utterances.
- A limited number of objective utterances hindered training the model to discriminate between subjective and objective utterances.
- The video only had one speaker. We did not include the chat shows and the discussions.

The data set was generated in Urdu and the processing only involved Urdu utterances without converting them to English.

VI. CONCLUSION AND FUTURE WORK

This paper presents a multimodal SA framework for the Urdu language, which detects the polarity of the sentences extracted from the videos. The videos were analyzed for visual, acoustic, and textual features. Unimodal and multimodal datasets were used in the experiments, and the videos themselves were used as data sources. Various topics and opinions were discussed in the videos. We also tried combining the modalities at the feature and decision levels. According to our results, feature and decision level integration improved the prediction performance.

Furthermore, the fact that Urdu is a resource-constrained language did not preclude us from conducting a useful analysis of the Urdu language using cutting-edge algorithms. Deep learning techniques that are used for feature extraction can also be used to extract useful features from a set of data, enabling a more accurate discovery of hidden patterns.

While deep learning algorithms are typically applied to large datasets, there are still numerous research studies in which the collection of large amounts of data is impossible due to time or other constraints. Researchers have proposed a variety of workarounds for this bottleneck, as they will frequently have only somewhat limited data to solve a problem in practice. In that case, experiments such as those described in the references [1]–[3] and [14] continue to be conducted and published. However, alternatives, such as the use of synthesized data or pre-trained networks, have been suggested by researchers, and we intend to investigate them in the future. One disadvantage of using a small number of observations is that it increases the likelihood of overfitting and producing inaccurate results. To some extent, we addressed this issue by combining early and late predictions to produce a more realistic outcome prediction. We now intend to expand our dataset to conduct future experiments.

In the future, we plan to accomplish the following goals. Our proposed deep learning modal can improve MSA results by adding different types of videos and data to the existing dataset. There are also other deep learning models that can be used on the dataset to see which one is the most accurate and gives the best performance. MSA in Urdu was a major focus of this paper, but other languages will be explored in the future. Unimodal datasets will be integrated using a variety of fusion techniques.

REFERENCES

[1] K. Dashtipour, M. Gogate, E. Cambria, and A. Hussain, "A novel context-aware multimodal framework for Persian sentiment analysis," *Neurocomputing*, vol. 457, pp. 377–388, Oct. 2021, doi: 10.1016/j.neucom.2021.02.020.

[2] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, Jul. 2019, doi: 10.1016/j.patrec.2019.04.024.

[3] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics," in *Proc. IEEE Symp. Comput. Intell. Human-Like Intell. (CIHLI)*, Apr. 2013, pp. 108–117, doi: 10.1109/CIHLI.2013.6613272.

[4] (2021). *A Guide to Urdu Language*. [Online]. Available: https://cudoo.com/blog/a-guide-to-the-urdu-language/

[5] A. Khattak, M. Z. Asghar, A. Saeed, I. A. Hameed, S. Asif Hassan, and S. Ahmad, "A survey on sentiment analysis in urdu: A resource-poor language," *Egyptian Informat. J.*, vol. 22, no. 1, pp. 53–74, Mar. 2021.

[6] M. Caon, L. Angelin. Yue, O. A. Khaled, and E. Mugellini, "Context-aware multimodal sharing of emotions," in *Proc. Int. Conf. Hum.-Comput. Interact.* (Lecture Notes in Computer Science). 2013, pp. 19–28, doi: 10.1007/978-3-642-39342-6_3.

[7] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5647–5657. [Online]. Available: https://www.researchgate.net/publication/336996537_Contextaware_Interactive_Attention_for_Multimodal_Sentiment_and_Emotion_Analysis

[8] N. Mukhtar and M. A. Khan, "Urdu sentiment analysis using supervised machine learning approach," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 2, Feb. 2018, Art. no. 1851001, doi: 10.1142/S0218001418510011.

[9] N. Mukhtar, M. A. Khan, and N. Chiragh, "Lexicon-based approach outperforms supervised machine learning approach for Urdu sentiment analysis in multiple domains," *Telematics Informat.*, vol. 35, no. 8, pp. 2173–2183, Dec. 2018, doi: 10.1016/j.tele.2018.08.003.

[10] K. Mehmood, D. Essam, K. Shafi, and M. K. Malik, "An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102368.

[11] H. Ghulam, F. Zeng, W. Li, and Y. Xiao, "Deep learning-based sentiment analysis for Roman Urdu text," *Proc. Comput. Sci.*, vol. 147, pp. 131–135, Apr. 2019.

[12] Z. Mahmood, I. Safder, R. M. A. Nawab, F. Bukhari, R. Nawaz, A. S. Alfakeeh, N. R. Aljohani, and S.-U. Hassan, "Deep sentiments in Roman Urdu text using recurrent convolutional neural network model," *Inf. Process. Manage.*, vol. 57, no. 4, Jul. 2020, Art. no. 102233.

[13] Q. Rajput, "Ontology based semantic annotation of Urdu language web documents," *Proc. Comput. Sci.*, vol. 35, pp. 662–670, Jul. 2014.

[14] V. P. Rosas, R. Mihalcea, and L. P. Morency, "Multimodal sentiment analysis of Spanish online videos," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 38–45, May/Jun. 2013, doi: 10.1109/MIS.2013.9.

[15] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[16] A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor, and A. Basit, "Extractive text summarization models for Urdu language," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102383.

[17] C. Gan, Q. Feng, and Z. Zhang, "Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis," *Future Gener. Comput. Syst.*, vol. 118, pp. 297–309, May 2021.

[18] I. Abu Farha and W. Magdy, "A comparative study of effective approaches for Arabic sentiment analysis," *Inf. Process. Manage.*, vol. 58, no. 2, Mar. 2021, Art. no. 102438.

[19] S. Smetanin and M. Komarov, "Deep transfer learning baselines for sentiment analysis in Russian," *Inf. Process. Manage.*, vol. 58, no. 3, May 2021, Art. no. 102484, doi: 10.1016/j.ipm.2020.102484.

[20] A. Kumar, K. Srinivasan, W.-H. Cheng, and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102141, doi: 10.1016/j.ipm.2019.102141.

[21] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data," *J. Big Data*, vol. 5, no. 1, pp. 1–18, Dec. 2018, doi: 10.1186/s40537-018-0120-0.

[22] H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 474–478, doi: 10.1109/ICSIPA.2017.8120658.

[23] A. Krishna, J. Zambreno, and S. Krishnan, "Polarity trend analysis of public sentiment on YouTube," in *Proc. 19th Int. Conf. Manage. Data (COMAD)*, 2013, pp. 125–128.

[24] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Sentiment analysis of Urdu language: Handling phrase-level negation," in *Proc. Mexican Int. Conf. Artif. Intell.*, vol. 7094, 2011, pp. 382–393, doi: 10.1007/978-3-642-25324-9_33.

[25] Y. Li, K. Zhang, J. Wang, and X. Gao, "A cognitive brain model for multimodal sentiment analysis based on attention neural networks," *Neurocomputing*, vol. 430, pp. 159–173, Mar. 2021.

[26] K. A. Araño, C. Orsenigo, M. Soto, and C. Vercellis, "Multimodal sentiment and emotion recognition in hyperbolic space," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115507, doi: 10.1016/j.eswa.2021.115507.

[27] S. Vashishtha and S. Susan, "Inferring sentiments from supervised classification of text and speech cues using fuzzy rules," *Proc. Comput. Sci.*, vol. 167, pp. 1370–1379, Mar. 2020, doi: 10.1016/j.procs.2020.03.348.

[28] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, and B. Wang, "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Inf. Fusion*, vol. 62, pp. 14–31, Oct. 2020, doi: 10.1016/j.inffus.2020.04.003.

[29] X. Li and M. Chen, "Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language," in *Chinese Computational Linguistics* (Lecture Notes in Computer Science), vol. 12522, M. Sun, S. Li, Y. Zhang, Y. Liu, S. He, and G. Rao, Eds. Cham, Switzerland: Springer, 2020, pp. 359–373, doi: 10.1007/978-3-030-63031-7_26.

[30] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal sentiment analysis via RNN variants," in *Proc. IEEE Int. Conf. Big Data, Cloud Comput., Data Sci. Eng. (BCD)*, May 2019, pp. 19–23, doi: 10.1109/BCD.2019.8885108.

[31] Y. Yao, V. Pérez-Rosas, M. Abouelenien, and M. Burzo, "MORSE: MultimOdal sentiment analysis for real-life settings," in *Proc. Int. Conf. Multimodal Interact.*, New York, NY, USA, Oct. 2020, pp. 387–396, doi: 10.1145/3382507.3418821.

[32] I. O. Hussien and Y. H. Jazyah, "Multimodal sentiment analysis: A comparison study," *J. Comput. Sci.*, vol. 14, no. 6, pp. 804–818, 2018, doi: 10.3844/jcssp.2018.804.818.

[33] M. A. Ullah, M. M. Islam, N. Binti Azman, and Z. M. Zaki, "An overview of multimodal sentiment analysis research: Opportunities and difficulties," in *Proc. IEEE Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, Feb. 2017, pp. 1–6, doi: 10.1109/ICIVPR.2017.7890858.

[34] Q. Portes, J. Carvalho, J. Pinquier, and F. Lerasle, "Multimodal neural network for sentiment analysis in embedded systems," in *Proc. 16th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, vol. 5, 2021, pp. 387–398, doi: 10.5220/0010224703870398.

[35] F. Ali, D. Kwak, P. Khan, S. El-Sappagh, A. Ali, S. Ullah, K. H. Kim, and K.-S. Kwak, "Transportation sentiment analysis using word embedding and ontology-based topic modeling," *Knowl.-Based Syst.*, vol. 174, pp. 27–42, Jun. 2019, doi: 10.1016/j.knosys.2019.02.033.

[36] S. Rahmani, S. Hosseini, R. Zall, M. Reza Kangavari, S. Kamran, and W. Hua, "Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects," 2021, *arXiv:2106.14174*. [Online]. Available: http://arxiv.org/abs/2106.14174

[37] R. Bibi, U. Qamar, M. Ansar, and A. Shaheen, "Sentiment analysis for Urdu news tweets using decision tree," in *Proc. IEEE 17th Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, May 2019, pp. 66–70.

[38] *The Editors of Encyclopaedia Britannica (5 December 2019), Urdu language, Encyclopaedia Britannica*, Sep. 2020.

UROOBA SEHAR received the bachelor's degree in software engineering from Riphah International University, in 2020, where she is currently pursuing the master's degree in software engineering.

SUMMRINA KANWAL received the Ph.D. degree from the University of Stirling. She is currently working as a Visiting Postdoctoral Fellow with the Cognitive Big Data and Cybersecurity Laboratory (CogBID), Napier University. She is also working as an Assistant Professor with the School of Computing and Informatics, Saudi Electronic University, Saudi Arabia.

KIA DASHTIPUR received the M.Sc. degree in advanced computer system development and the Ph.D. degree in computing science from the University of Stirling, in 2014 and 2019, respectively. From 2018 to 2019, he was a Postdoctoral Research Associate with Edinburgh Napier University. He is currently a Research Associate with the James Watt School of Engineering, University of Glasgow. His main research interests include natural language processing, machine learning, and speech enhancement.

USAMA MIR (Senior Member, IEEE) received the B.S. degree (Hons.) in computer engineering from the Balochistan University of IT, Engineering and Management Sciences, Pakistan, in 2006, and the master's and Ph.D. degrees in computer science from the Troyes University of Technology, France, in 2008 and 2011, respectively. From 2011 to 2012, he was a Postdoctoral Fellow with Telecom Bretagne, France. From 2012 to 2015, he was the Head of the Electronics Engineering Department, Iqra University, Islamabad, Pakistan. He is currently an Associate Professor with the Department of Computing and IT, Saudi Electronic University, Saudi Arabia. His research interests include big data analysis, energy management, blockchains, MIMO technology, resource allocation, and handoff management in cognitive radio systems, digital currencies, wireless communications and networking, and multi-agent systems. He is an Associate Editor of IEEE Access.

UBAID ABBASI received the M.S. degree from SUPELEC, Rennes, France, in 2008, and the Ph.D. degree from the University of Bordeaux, France, in 2012. He also worked as a Senior Research Fellow at the University of Quebec, Montreal, QC, Canada, working on a project funded by Ericsson Canada. He is currently an Assistant Professor with the Department of Sciences, GPRC, Grande Prairie, AB, Canada. His research interests include inter-container communications, energy management, data center communication issues, device-to-device communication in next-generation 5G networks, wireless communications, and big data analysis.

FAIZA KHAN received the master's degree in software engineering from Riphah International University, Islamabad, in September 2019. Her research interests include machine learning and evolutionary computation.