BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Using Deep Learning Models for Multimodal Sentence Level Sentiment Analysis of Sign Language

Osondu Oguike * , ***Mpho Primus***

Institute for Intelligent Systems, University of Johannesburg, Auckland Park, Johannesburg 2092, South Africa

ABSTRACT

Deaf and hard-of-hearing individuals communicate with signs such as hand signals, gestures, facial expressions, and body movements. This medium of communication is called sign language, which is a non-verbal, visual means of communication. However, some non-deaf and non-hard-of-hearing individuals do not understand sign language; those who understand it use it to communicate with deaf and hard-of-hearing individuals. Some active users of social media are deaf and hard-of-hearing individuals; therefore, it is necessary to develop technological tools that will guarantee effective communication between deaf & hard-of-hearing individuals and non-deaf & non-hard-of-hearing individuals, especially across various social media platforms. Sentiment analysis of sign language is one such technological tool that helps to communicate the polarity expressed in sign language. A multimodal approach to sentiment analysis of sign language is the focus of this study, which uses a multimodal sign language dataset to train two Deep Learning models. The dataset consists of video clips of sentence-level sign language and textual equivalents. The dataset trains a deep convolutional neural network model called VGG16 for visual modality. The other Deep Learning model, which the dataset trains, is Bidirectional Encoder Representation from Transformer, BERT for textual modality. The results of the performance metrics showed that the multimodal approach performed better than the single-modality text-based approach.

Keywords: Sign Language; Sentiment Analysis; Visual Signal; Textual Modality; Multimodality; Sign Language Sentiment Recognition

*CORRESPONDING AUTHOR:

Osondu Oguike, Institute for Intelligent Systems, University of Johannesburg, Auckland Park, Johannesburg 2092, South Africa; Email: osonduo@uj.ac.za or mraborife@uj.ac.za

ARTICLE INFO

Received: 10 March 2025 | Revised: 18 March 2025 | Accepted: 6 April 2025 | Published Online: 16 April 2025
DOI: https://doi.org/10.30564/fls.v7i4.9041

CITATION

Oguike, O., Primus, M., 2025. Using Deep Learning Models for Multimodal Sentence Level Sentiment Analysis of Sign Language. Forum for Linguistic Studies. 7(4): 935–948. DOI: https://doi.org/10.30564/fls.v7i4.9041

1. Introduction

Sentiment analysis identifies the polarity (positive, neutral, or negative) or emotional states (anger, sadness, happiness, disgust, fear, surprise, etc.) expressed in written text, spoken words, or visual expressions. While written text and spoken words are typically in a specific natural language, visual expression, especially in sign language, is a medium of communication exclusively used by the deaf and hard-of-hearing individuals, relying on hand gestures and facial expressions.

Deaf and hard-of-hearing individuals are integral to our society. According to the World Health Organization (WHO), they constitute about 5 percent of the world population[1]. Therefore, developing automated tools that will help in communication between the deaf & hard-of-hearing individuals and non-deaf & non-hard-of-hearing individuals has become very necessary[2].

Like various natural language processing tasks, various automated tasks can be performed in sign language. These include sign language recognition[3], sign language translation[4], and sign language sentiment analysis/emotion recognition[5].

Gloss in sign language is the basic unit of sign language information. Therefore, complete words will be composed of several glosses[6]. A task involving a single gloss in a sign language presentation video is called Isolated Sign Language Recognition (ISLR), whereas Continuous Sign Language Recognition (CSLR) involves multiple glosses in a sign language presentation video[6]. To facilitate communication between signers and non-signers, sign language must be converted into an equivalent natural language. Sign language recognition transforms sign language into the textual modality of the equivalent natural language, while sign language translation converts it into the audio or text modalities, and vice versa[5, 7].

After converting sign language into its natural language equivalent in either textual or audio form, non-signers can understand what signers are communicating. However, recognizing the sentiment of sign language is another essential task. Most studies have recognized sentiment or emotion in sign language primarily through facial expressions[8]. Moreover, adopting a multimodal approach to sentiment understanding, which considers both the recognized sign language (text) and visual expressions (facial expressions), is a more effective approach to sentiment understanding because sentiment information is available in both modalities.

The significant contributions of this study are twofold; firstly, this study transforms an existing sentence-level sign language translation and recognition dataset, created by Elakkiya and Natarajan[9], into a multimodal sign language sentiment recognition dataset, which includes both visual and textual modalities. The transformation process involved adding a new metadata, "Sentiment," which was labeled and used to perform multimodal sentence-level sentiment analysis of sign language. The video of the sign language was used for the visual modality, while the "Sentences" metadata was used for the textual modality. The class label of the new sentence-level multimodal sign language dataset is the Sentiment metadata. Secondly, this study performs multimodal sentiment analysis of sign language using textual and visual modalities, which have not been done in previous studies.

1.1. Problems Statement

This study solves the following problems:

- Signers and non-signers are unable to communicate
  Signers and non-signers are not able to communicate with each other because they cannot understand each other. As a result, non-signers need a human interpreter or an automated tool, which this study provides. This will help to understand the thoughts, opinions, and sentiments expressed by singers.
- Low-resourced sign languages
  Sign language can be considered a low-resourced language due to the unavailability of technological resources for sign language. Therefore, developing technological tools for sentence-level sentiment analysis of sign language is a step toward making it a high-resourced language.
- Signer's digital divide
  Without technological tools like sentence-level sentiment analysis for sign language, signers are left behind in global online communication, via different social media platforms. However, the existence of a sentiment analysis tool for sign language promotes inclusion, as it enables signers to be easily understood by non-signers, fostering better global communication.
- Information loss in other modalities

Like natural languages, information in sentence-level sign language can be embedded in multiple modalities. Therefore, considering only one modality may result in the loss of information present in other modalities.

- Sign language faces the danger of extinction
  The absence of sentence-level sentiment analysis for sign language discourages non-signers from learning it, which can lead to sign language extinction due to a declining number of learners. Therefore, if the number of people using sign language for communication is declining, sign language faces the risk of extinction.

1.2. The Aim and Objectives of the Study

The study aims to perform a multimodal sentence-level sentiment analysis of sign language with the following objectives:

- Download the sentence-level sign language dataset, ISL-CSLTR, which was created for sign language translation and recognition.
- Transform the dataset into a multimodal sentiment analysis dataset by adding the metadata “Sentiment” and using the natural language processing automated tool, TextBlob, to label it and afterwards validate the sentiment label manually.
- Use the modified sign language dataset to train and validate two Deep Learning models, which are a deep convolutional neural network model (VGG-16) for the visual modality of the dataset and a BERT language model for the textual modality of the dataset.
- Use a decision-level (feature-level) fusion technique to combine the results of the training and validation of the two models.
- Evaluate the performance of the training and validation using various performance metrics.

1.3. Significance of the Study

Based on the problem that this study solves, together with its aim and objectives, the importance of this study includes the following: bridges the communication gap between signers and non-signers; makes sign language a high-resourced sign language[10]; signers inclusiveness[11]; effective sign language sentiment analysis through modality inclusion[5]; preservation of sign language from extinction[12].

2. Materials and Methods

2.1. Overview of Sign Language Studies

Globally, there are about 135 different sign languages, each country or region having its own[5]. The available sign languages include the following: American Sign Language, Indian Sign Language, Chinese Sign Language, etc[3]. Each sign language has its own set of signs and gestures for different alphabets, words, and sentences. The primary components of sign language are hand gestures and facial expressions[5].

Two classes of signals in sign language exist, which are manual signals and non-manual signals. Manual signals use hand shape, position, location, and movement, while non-manual signals use other parts of the body, such as eye gaze and movement, lip patterns, mouthing, body movement, and orientation[5]. Furthermore, manual signals are used for words, phrases, and sentences, while non-manual signals are used to demonstrate grammatical and emotional information[13]. Information that is provided by non-manual signals helps to complement manual signals; this allows for comprehensive and accurate messages in sign language[5, 14]. Manual and non-signals provide information, which are in different modalities, therefore, the combined manual and non-manual signals form a multimodal approach to sentence-level sign language.

Different studies on sign language, as revealed in the literature fall into three main categories, which are: sign language recognition, sign language translation, and sign language datasets[5]. Sign language recognition, as shown in **Figure 1**, can further be classified into hand gesture recognition, facial recognition, and combined recognition. Sign language recognition has been studied from four angles: background, gesture, special hardware utilization, and continuity[5].

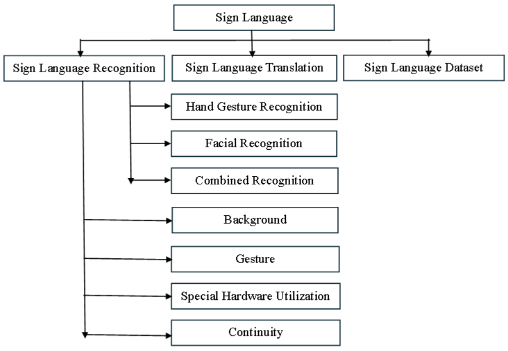


Figure 1. An Overview of Sign Language Studies.

Sign language recognition and translation determine the natural language equivalent of sign language in textual and audio modalities, respectively, which have been extensively researched [1, 15–17]. Furthermore, some datasets for sign language recognition and translation are available [9, 18]. However, few studies have focused on understanding the sentiments and emotions in sign languages [2]. This informs the motivation for this study.

2.2. Sign Language Recognition

The sign language processing task that identifies the textual equivalent of sign language at the word or sentence level is called sign language recognition. It transforms a piece of sign language into text [6]. The device used for the sign language determines the approach that this sign language processing task uses [3]. The approaches are the vision-based approach, which uses images or videos that a video camera captures, and the glove-based approach, which uses data gloves, a special device that captures hand poses and movements [3].

Different studies have used different approaches for this task; some have used the vision-based approach, such as Mariappan et al. (2019) and Kamal et al. (2017) [19, 20], while others have used the data gloves approach, such as Abhishek et al. (2016) and Xiao et al. (2019) [21, 22]. Data gloves contain many sensors, and they are worn as hand gloves on both hands. Though the data glove approach has high speed and high precision advantages due to its large number of sensors [23], it is expensive and may not be within the reach of the average deaf individual [3]. Furthermore, its complexity has made its usage limited, compared to the vision-based approach [6].

Both the vision-based and data gloves approaches consist of the following stages: image acquisition, image preprocessing, feature extraction, sign classification, and sign translation [3, 6, 19].

Different machine learning and Deep Learning models have been used for sign language recognition, together with sign language recognition datasets [20]. For example, SVM has been used to recognize Chinese sign language alphabets and isolated words [24], while CNNs have been used to recognize Chinese sign language isolated words and continuous sentences [25, 26]. Furthermore, Transformers, like Bidirectional Encoder Representation from Transformer (BERT), have been used for sign language recognition [6, 27].

Sign language recognition datasets are in different forms, shapes, and sizes. These datasets can be used to recognize alphabets, words, or continuous sentences in different sign languages. Some datasets are for alphabet recognition, which are called fingerspelling datasets, such as the American fingerspelling dataset [28]. Other datasets combine signs of single words and continuous sentences for specific countries, for example, the Danish sign language dataset by and the German sign language dataset [29, 30]. All sign language recognition datasets contain video clips of signers.

The challenges of sign language recognition are signer-independent conditions, short length of videos, lack of consideration for unseen sentences, multi-signer condition, inability to recognize sign language outside the dataset, high model complexity, and lack of online recognition [6].

2.3. Sign Language Translation

The transformation of sign language into either spoken language or written text, or vice versa, or both is called sign language translation task [5, 7]. It consists of the following processes: mapping sign language to spoken language text, mapping spoken language text to sign language, mapping speech in each language to its equivalent sign language, and mapping sign language to speech [7]. While sign language recognition converts sign language to text, sign language translation provides a broader task by converting sign language to text or speech and vice versa. The use of neural networks, generative models, and custom datasets is required for sign language translation tasks [5].

However, sign language production has been used in the literature as a task that involves mapping text or speech in a natural language to sign language [31]. This means that sign language production is part of sign language translation, but sign language production is not part of sign language recognition.

Different sign language translation systems have been developed; they include the following: a mobile platform that translates American Sign Language into speech [32], an automatic speech/text-to-sign language system that first converts speech to text and then translates the text to Arabian sign language [33].

2.4. Sign Language Sentiment Analysis/Emotion Recognition

Facial expression features, like eye gaze, eyebrows, eye blinks, and mouth movements, can be used to communicate human emotions and sentiments. These facial expression features help to determine the emotion and sentiment expressed in sign language[34]. Grammatical Facial Expressions (GFEs) are a group of facial expressions, that sign language uses to support grammatical construction and elimination of sign ambiguity in sign language[35].

Deep Learning models have shown more superiority in recognizing facial expressions and determining the emotions or sentiments expressed by the singer than traditional machine learning—for example, multi-region ensemble CNNs, CNN-Recurrent Neural Networks (CNN-RNN), and deep CNN methods for learning from noisy labels, using facial expression recognition[8, 36, 37]. Similarly, this study has trained VGG16, a deep CNN, for the visual modality of a sign language dataset.

The facial expression recognition model is a joint framework, which consists of two models that are used for capturing temporal appearance features from image sequence, and for extracting temporal geometry features from facial landmark points[38].

A comprehensive survey of facial attribute analysis revealed two Deep Learning-based issues that are related to facial attributes: Facial Attribute Estimation (FAE) and Facial Attribute Manipulation (FAM)[39]. The Facial Attribute Estimation (FAE) was used to verify if a particular facial attribute is present in an image, while Facial Attribute Manipulation (FAM) was used to synthesize or remove a particular facial attribute.

Furthermore, Kinect has been used to propose an emotion recognition system, which uses a 3D point cloud to remove non-facial points and surface curvature features to recognize three types of emotions[40]. When tested with a support vector machine, it achieved an accuracy of 77.4%, with about 707 image segments.

2.5. Multimodal Approach to Sign Language Sentiment Understanding

A multimodal approach can be used to enhance the understanding of sentiment and emotion expressed in sign language. This involves a combination of manual and non-manual signals[33]. However, few studies have considered the multimodal approach to sentiment understanding in sign language[14]; they include the following:

- Hidden Markov Model (HMM) was used to develop a multimodal system that simultaneously recognizes hand gestures and head movements[41].
- Hierarchical Conditional Random Fields (H-CRF) and Support Vector Machines (SVM) were used as a framework to identify hand gestures and facial expressions[42].
- Three cameras were used to capture different orientations for a multimodal framework[43].
- Global hand locations/motions and local hand gesture details with a Hierarchical Attention Network and Latent Space (LS-HAN) were used for a multimodal framework of continuous sign language recognition[44].
- Manual and non-manual signs were used to combine a sequential belief network for sign language recognition[45].
- A combination of facial expressions captured by Kinect and hand gestures captured by Leap Motion with a public dataset of Indian sign language was used to train HMM and Bayesian classification[35].

Accurate sign language recognition requires a multimodal approach that combines facial recognition with hand gesture recognition. Similarly, accurate emotion and sentiment recognition in sign language require a multimodal approach that combines the visual components of sign language with the recognized textual equivalent for sentiment understanding.

Based on evidence from the reviewed literature, we did not see any study that combined the videos of sign language and the equivalent text to conduct multimodal sentiment analysis of sign language; this is where this study advances previous research.

2.6. Methods

The methods used for the multimodal sentence-level sentiment analysis of sign language include the following: a method for accessing and modifying the ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition dataset, from a sign language recognition and translation dataset to multimodal sign language sentiment recognition dataset. Another method used in this study is the method for the integration/fusion of the different modalities of the dataset.

2.6.1. Dataset Description

The ISL-CSLTR dataset was created for Indian sign language translation and recognition tasks, which is available in Mendeley public repository [13]. However, as part of the contribution of this study, we modified the dataset so that it could be used for multimodal sentiment analysis. The modification of the dataset involved adding metadata, Sentiment, which was annotated and labeled using an automated tool, TextBlob, together with the text of each sentence-level sign language provided in the Sentences metadata of the dataset. The annotation and labeling of the Sentiment metadata helped to determine the sentiment of each sentence-level sign language video clip of the ISL-CSLTR dataset; this was done manually with the help of two signers. To reduce bias and errors from the labelling of the automated tool, TextBlob, manual validation of the labelling of the TextBlob was done. This validation involves two human annotators, who independently determined the sentiment of the recognized text. The final validated sentiment was obtained using the technique of majority voting of the labels of TextBlob and the two independent annotators. The modified CSV file of the dataset, which was used for training, has the following metadata/attributes: SN, Sentences, Filename, and Sentiment. The description of each of these metadata of the dataset are as follows: SN is the serial number for each of the instances of sentence level sign language dataset, Sentences is the text equivalent of the sentence level sign language, Filename is the name of the MP4 video clip for the sentence level sign language, and Sentiment is the combined sentiment polarity of the text and video of the sentence-level sign language dataset. Initially, this CSV file contains a total of 492 instances of data; however, after all the sentence-level sign language video clips were downloaded, the frames/images of the sign language video clips generated from the sign language video clips using the Python library, CV2, the dataset was expanded to 4,193 instances of data. The distribution of the sentiment polarities of the images/frames is shown in **Figure 2**, which shows that the dataset is not biased, based on the distribution of the dataset. This was used to train visual and textual modalities. Part of the preprocessing task done on the dataset before the training involved ensuring that the relevant metadata like Sentiment and Sentences were not blank.

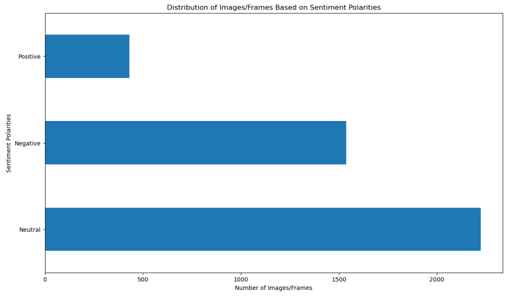


Figure 2. Distribution of Images/Frames based on Sentiment Polarities.

2.6.2. Fusion Techniques

Fusion techniques describe how the different modalities, such as textual and visual modalities, are combined. Though there are different techniques for fusion, in this study, we have used the late/decision-level feature technique. This is because of the simplicity and ease of use of the late/decision-level fusion technique [46].

The late/decision-level fusion technique independently extracts the features of each modality of the dataset, trains an appropriate model for that modality, and combines the results of the training of the different modalities to obtain the prediction result. **Figure 3** illustrates this fusion technique, which is our proposed framework for multimodal sentence-level sentiment analysis of sign language.

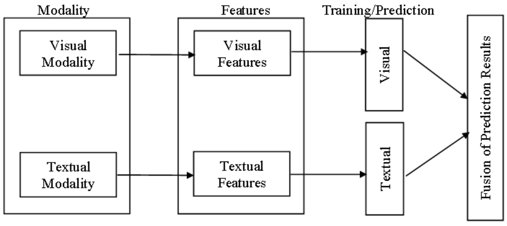


Figure 3. Late/Decision Level Fusion Technique.

2.7. Model Training and Validations

After pre-processing the dataset, it was split into training and validation sets in the ratio of 4:1. Afterwards, the VGG-16 and BERT models were trained independently for the visual and textual modalities, respectively. Deep Learning models were chosen over other traditional machine learning models based on high performance and ability to handle large and unstructured datasets, like texts and images [47].

The results of this independent training were combined using the late/decision-level fusion technique. The method of averaging was used in the late/decision level fusion technique to combine the results of the independent training of the two models.

2.8. Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used copy editing tools in ChatGPT and Grammarly to improve language and readability. This was used in the following manner: after writing the paper, different sections of the paper were copied and used to ask ChatGPT to fix and correct grammatical errors in the text. Similarly, Grammarly was used for the same purpose of copy editing. After using these tools, the author(s) reviewed and edited the content as needed and took full responsibility for the content of the publication.

3. Results

The training and validation datasets were split in the ratio of 4:1, respectively; after the training and validation for the visual modality, the following performance metrics were recorded for 50 epochs: Loss, Accuracy, Val_Loss, and Val_Accuracy. **Table 1**, **Table 2**, and **Table 3** show the results recorded for the visual modality, textual modality, and multimodality, respectively. The result of the multimodality, which was shown in **Table 3,** was obtained using the average of the results for the visual and textual modalities. Furthermore, the visualization of the various performance metrics results for visual, textual, and multimodality are shown in **Figure 4**, **Figure 5**, and **Figure 6**, respectively.

Table 1. Results of the Performance Metrics for Visual Modality.

		Visual Modality		
Epoch	**Loss**	**Accuracy**	**Val_Loss**	**Val_Accuracy**
1	0.0987	0.9645	0.0094	0.9976
2	0.0887	0.9684	1.81E−02	0.994
n	0.0793	0.9711	1.05E−02	0.9988
4	7.66E−02	0.9758	1.54E−02	0.9964
5	6.46E−02	0.9773	1.39E−02	0.9964
6	1.06E−01	0.9651	2.92E−02	0.9869
7	1.58E−01	0.9359	3.37E−02	0.9952
8	1.25E−01	0.9463	1.45E−02	0.9988
9	1.10E−01	0.9648	2.36E−02	0.994
10	1.38E−01	0.9439	7.90E−03	0.9988
11	1.38E−01	0.9377	1.16E−02	0.9864
12	1.38E−01	0.9332	2.57E−02	0.9869
13	1.20E−01	0.9478	4.50E−03	1
14	1.33E−01	0.9392	9.40E−03	0.9976
15	1.53E−01	0.9231	8.30E−03	0.9988
16	1.44E−01	0.9323	4.10E−03	1
17	1.25E−01	0.9377	1.30E−02	0.9952
18	8.94E−02	0.9597	3.40E−03	0.9988
19	9.66E−02	0.9559	1.16E−02	0.9964
20	1.32E−01	0.9422	2.19E−02	0.994
21	1.29E−01	0.941	3.20E−03	0.9988
22	1.88E−01	0.9067	1.45E−02	0.9964
23	1.57E−01	0.9237	3.60E−03	1
24	1.44E−01	0.9281	6.34E−02	0.9702
25	1.43E−01	0.9275	3.67E−02	0.9869
26	1.28E−01	0.9359	3.10E−02	0.9928
27	1.35E−01	0.9317	4.50E−03	0.9988
28	1.53E−01	0.9293	7.89E−02	0.9702
29	1.31E−01	0.9317	1.48E−01	0.9559

Table 1. *Cont.*

		Visual Modality		
Epoch	**Loss**	**Accuracy**	**Val_Loss**	**Val_Accuracy**
30	1.42E−01	0.9273	2.01E−02	0.9928
31	1.50E−01	0.9237	5.29E−01	0.8153
32	1.71E−01	0.9165	6.44E−02	0.9821
33	1.44E−01	0.9335	4.04E−02	0.9821
34	1.51E−01	0.9329	4.64E−02	0.9785
35	1.45E−01	0.9293	2.60E−02	0.9905
36	1.48E−01	0.9299	1.36E−01	0.9654
37	2.13E−01	0.9031	7.22E−02	0.9714
38	1.43E−01	0.9431	1.27E−02	0.9964
39	1.46E−01	0.935	3.61E−02	0.9833
40	1.47E−01	0.9383	4.10E−03	1
41	2.14E−01	0.9267	1.24E−01	0.9583
42	1.49E−01	0.946	1.03E−01	0.9619
43	1.41E−01	0.9475	5.44E−02	0.9833
44	1.31E−01	0.9547	1.85E−02	0.9952
45	1.30E−01	0.9499	5.22E−02	0.9809
46	1.38E−01	0.9502	1.92E−02	0.994
47	1.24E−01	0.9547	1.08E−01	0.969
48	1.28E−01	0.9547	5.96E−02	0.9821
49	1.50E−01	0.9487	2.36E−02	0.9905
50	1.22E−01	0.958	2.70E−03	0.9988
Average	0.134912	0.941624	0.044498	0.985056

Table 2. Results of the Performance Metrics for Textual Modality.

		Textual Modality		
Epoch	**Loss**	**Accuracy**	**Val_Loss**	**Val_Accuracy**
1	1.0655	04858	0.8574	0.6445
2	0.9537	05295	0.8869	0.6445
3	0.9443	0.528	0.8758	0.6445
4	0.9446	0.5278	0.8494	0.6445
5	0.9478	0.5249	0.8816	0.6445
6	0.9426	0.5295	0.8596	0.6445
7	0.9416	0.5297	0.8792	0.6445
8	0.9436	0.5297	0.8945	0.6445
9	0.9468	0.5297	0.8716	0.6445
10	0.9415	0.5306	0.8696	0.6445
11	0.9434	0.5295	0.8865	0.6445
12	0.9423	0.5302	0.8706	0.6445
13	0.9421	0.5297	0.8646	0.6445
14	0.9442	0 5304	0.8515	0.6445
15	0.9433	0.5302	0.8741	0.6445
16	0.9424	05304	0.8873	0.6445
17	0.9409	0.5316	0.8827	0.6445
18	0.9432	0.5283	0.8675	0.6445
19	0.9412	0.5297	0.8612	0.6445
20	0.9403	0 5273	0.8545	0.6445
21	0.9417	0.5302	0.8602	0.6445
22	0.9412	0.5302	0.8546	0.6445
23	0.9412	0.5302	0.8634	0.6445
24	0.9426	0.5302	0.8837	0.6445

Table 2. *Cont.*

		Textual Modality		
Epoch	**Loss**	**Accuracy**	**Val_Loss**	**Val_Accuracy**
25	0.9422	0.5302	0.8685	0.6445
26	0.941	0.5302	0.848	0.6445
27	0.9406	0.5304	0.8744	0.6445
28	0.9404	0.5304	0.8652	0.6445
29	0.9415	0.529	0.8616	0.6445
30	0.9415	0.5297	0.864	0.6445
31	0.9418	0.5304	0.8733	0.6445
32	0.9403	0 5304	0.8757	0.6445
33	0.9414	0.5304	0.8784	0.6445
34	0.9407	0.5304	0.8753	0.6445
35	0.9415	0 5304	0.8616	0.6445
36	0.9418	0.5295	0.8616	0.6445
37	0.9414	0.5304	0.8663	0.6445
38	0.94	0.5304	0.8628	0.6445
39	0.9404	0.5304	0.8643	0.6445
40	0.941	0.5304	0.8603	0.6445
41	0.9401	0.5302	0.8632	0.6445
42	0.941	0.5302	0.8642	0.6445
43	0.9422	0.5309	0.8533	0.6445
44	0.9404	0.5304	0.873	0.6445
45	0.9422	0.5304	0.8719	0.6445
46	0.9388	0.5304	0.8619	0.6445
47	0.9405	0.5299	0.8652	0.6445
48	0.9411	0.5304	0.863	0.6445
49	0.9392	0.5306	0.8557	0.6445
50	0.9398	0.5304	0.8733	0.6445
Average	0.944496	0.529002	0.86788	0.6445

Table 3. Results of the Performance Metrics for the Multimodality.

		Multimodality (Visual and Textual)		
Epoch	**Loss**	**Accuracy**	**Val_Loss**	**Val_Accuracy**
1	0.5821	0.72515	0.4334	0.82105
2	0.5212	0.74895	0.4525	0.81925
3	0.5118	0.74955	0.44315	0.82165
4	0.5106	0.7518	0.4324	0.82045
5	0.5062	0.7511	0.44775	0.82045
6	0.52405	0.7473	0.4444	0.8157
7	0.55	0.7328	0.45645	0.81985
8	0.5343	0.738	0.4545	0.82165
9	0.52835	0.74725	0.4476	0.81925
10	0.5396	0.73725	0.43875	0.82165
11	0.54055	0.7336	0.44905	0.81545
12	0.54005	0.7317	0.44815	0.8157
13	0.5309	0.73875	0.43455	0.82225
14	0.53875	0.7348	0.43045	0.82105
15	0.5479	0.72665	0.4412	0.82165
16	0.543	0.73135	0.4457	0.82225
17	0.5329	0.73465	0.44785	0.81985
18	0.5163	0.744	0.43545	0.82165
19	0.5189	0.7428	0.4364	0.82045

Table 3. *Cont.*

Multimodality (Visual and Textual)				
Epoch	**Loss**	**Accuracy**	**Val_Loss**	**Val_Accuracy**
201	0.53605	0.73475	0.4382	0.81925
21	0.53515	0.7356	0.4317	0.82165
22	0.56465	0.71845	0.43455	0.82045
23	0.54895	0.72695	0.4335	0.82225
24	0.54335	0.72915	0.47355	0.80735
25	0.5426	0.72885	0.4526	0.8157
26	0.5346	0.73305	0.4395	0.81865
27	0.5377	0.73105	0.43945	0.82165
28	0.5466	0.72985	0.47205	0.80735
29	0.536	0.73035	0.50485	0.8002
30	0.54155	0.7285	0.44205	0.81865
31	0.54575	0.72705	0.7009	0.7299
32	0.55565	0.72345	0.47005	0.8133
33	0.5428	0.73195	0.4594	0.8133
34	0.5457	0.73165	0.46085	0.8115
35	0.54335	0.72985	0.4438	0.8175
36	0.5449	0.7297	0.49885	0.80495
37	0.5774	0.71675	0.46925	0.80795
38	0.54145	0.73675	0.43775	0.82045
39	0.54335	0.7327	0.4502	0.8139
40	0.5438	0.73435	0.4322	0.82225
41	0.57705	0.72845	0.49355	0.8014
42	0.5449	0.7381	0.48335	0.8032
43	0.54145	0.7392	0.45385	0.8139
44	0.53585	0.74255	0.44575	0.81985
45	0.5362	0.74015	0.46205	0.8127
46	0.53855	0.7403	0.44055	0.81925
47	0.5323	0.7423	0.48645	0.80675
48	0.53435	0.74255	0.4613	0.8133
49	0.5447	0.73965	0.43965	0.8175
50	0.53105	0.7442	0.438	0.82165
Average	0.539704	0.735313	0.456189	0.814778

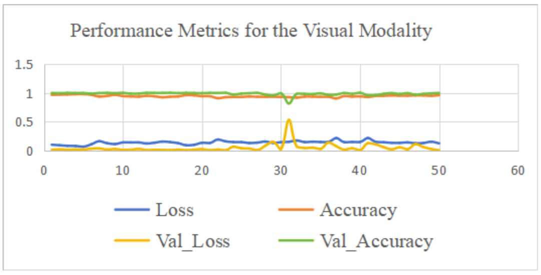


Figure 4. Visualization of the Results of the Performance Metrics for Visual Modality.

4. Discussion

Though the values of all the performance metrics range between 0 and 1, for optimal performance, the Loss and Val_Loss performance metrics should be close to 0, while Accuracy and Val_Accuracy performance metrics should be close to 1. Based on the chosen models for the two modalities, **Table 1** and **Table 2** show the results for all the performance metrics for the visual and textual modalities, respectively. The results show that the averages of the performance metrics for the visual modality are better than the averages of the performance metrics for the textual modality. Furthermore, from **Tables 2** and **3**, the results for all the averages of the performance metrics for the multimodality perform better than the averages for the performance metrics of the textual modality. These results show that visual-based sentence-level sentiment analysis of sign language performs better than text-based sentence-level sentiment analysis of sign language. These results are the same as results obtained in similar studies, which show better performance metrics, like F1 score and AUC score, for multimodal classification studies than unimodal classification studies [48, 49]. Furthermore, multimodal sentence-level sentiment analysis of sign language performs better than unimodal text-based sentence-level sentiment analysis of sign language [34]. For the BertforSequenceClassification pre-trained model, the confusion matrix and classification report are shown below in **Figures 7** and **8**.

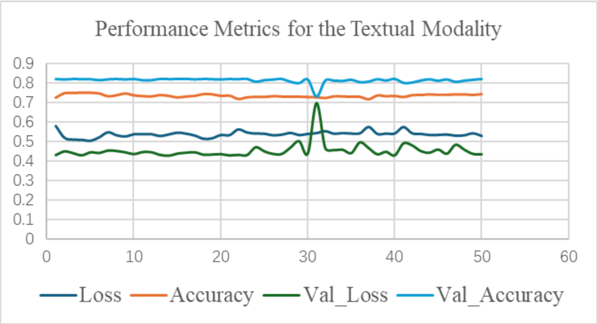


Figure 5. Visualization of the Results of the Performance Metrics for Textual Modality.

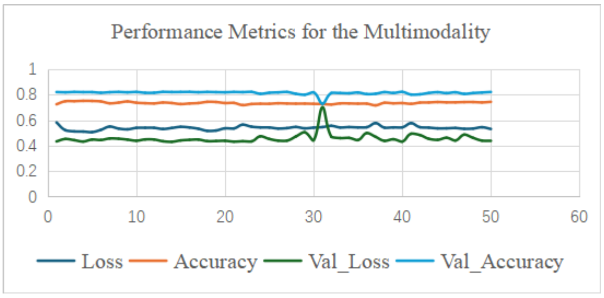
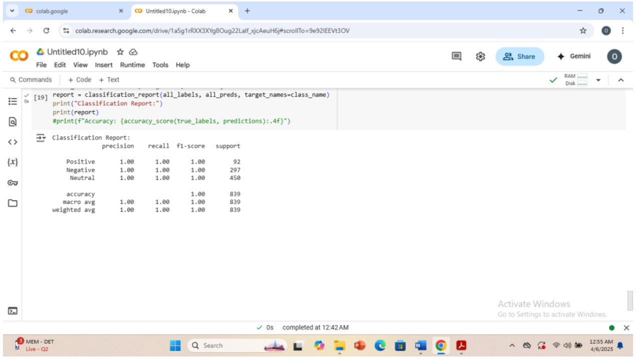


Figure 6. Visualization of the Results of the Performance Metrics for the Multimodality.



```
[19] report = classification_report(all_labels, all_preds, target_names=class_name)
     print("Classification Report:")
     print(report)
     #print(f"Accuracy: {accuracy_score(true_labels, predictions):.4f}")
```

```
Classification Report:
              precision    recall  f1-score   support

    Positive       1.00      1.00      1.00        92
    Negative       1.00      1.00      1.00       297
     Neutral       1.00      1.00      1.00       450

    accuracy                           1.00       839
   macro avg       1.00      1.00      1.00       839
weighted avg       1.00      1.00      1.00       839
```

Figure 7. Confusion Matrix for the BertforSequenceClassification Pre-Trained Model.

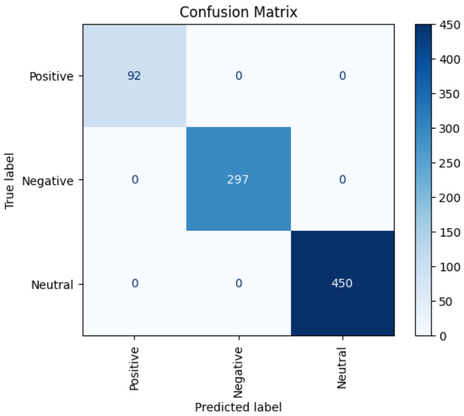


Figure 8. Classification Report for the BertforSequenceClassification Pre-Trained Model.

5. Conclusions

This study conducted a multimodal, sentence-level sentiment analysis of sign language by modifying an existing dataset and leveraging both visual and textual modalities. Using the averaging technique in the late fusion method, the study determined the predicted sentiments. The results confirm that, although the visual modality outperforms the textual modality individually, combining both in a multimodal approach yields better sentiment analysis than relying on the textual modality alone. This outcome highlights that sign language is inherently a visual means of communication, but integrating visual components with translated text offers more accurate sentiment information than using the translated text alone.

Author Contributions

The research article was written by the two authors, with the following contributions by the authors:Conceptualization, O.O. and M.P.; methodology, O.O.; software, O.O.; validation, M.P.; formal analysis, O.O.; investigation, O.O.; resources, M.P.; data curation, O.O.; writing—original draft preparation, O.O.; writing—review and editing, M.P.; visualization, O.O.; supervision, M.P.; project administration, M.P.; funding acquisition, M.P. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by [University of Johannesburg] transaction number [9BU5759115782932V].

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The dataset used for this study was downloaded from the Mendeley public repository, available at: https://data.mendeley.com/datasets/kcmpdxky7p/1 Published: 22 January 2021 | Version 1 | DOI: 10.17632/kcmpdxky7p.1.

Acknowledgments

During the preparation of this work, O. E. Oguike acknowledges the use of copy-editing tools in ChatGPT and Grammarly to improve language and readability. After using ChatGPT, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Akandeh, A., 2022. Sentence-Level Sign Language Recognition Framework. Proceedings of the 2022 International Conference on Computational Science and Computational Intelligence (CSCI); 14–16 December 2022; Las Vegas, NV, USA. pp. 1436–1441. DOI: https://doi.org/10.1109/CSCI58124.2022.00256

[2] Jamwal, A., Vasukidevi, G., Malleswari, T.Y.J.N., et al., 2022. Real Time Conversion of American Sign Language to text with Emotion using Machine Learning. Proceedings of the Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC-2022); 10–12 November 2022; Dharan, Nepal. pp. 603–609.

[3] Nimisha, K.P., Jacob, A., 2020. A Brief Review of the Recent Trends in Sign Language Recognition. Proceedings of the 2020 International Conference on Communication and Signal Processing; 28–30 July 2020; Chennai, India. pp. 0186–0190.

[4] Sharma, A., Panda, S., Verma, S., 2020. Sign Language to Speech Translation. Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 01–03 July 2020; Kharagpur, India. pp. 1–8.

[5] Alaghband, M., Maghroor, H.R., Garibay, I., 2023. A survey on sign language literature. Machine Learning with Applications. 14, 100504. DOI: https://doi.org/10.1016/j.mlwa.2023.100504

[6] Tao, T., Zhao, Y., Zhu, J., 2024. Sign Language Recognition: A Comprehensive Review of Traditional and Deep Learning Approaches, Datasets, and Challenges. IEEE Access. 12, 75034–75060. DOI: https://doi.org/10.1109/ACCESS.2024.3398806

[7] Núñez-Marcos, A., Perez-de-Viñaspre, O., Labaka, G., 2023. A survey on Sign Language machine translation. Expert Systems with Applications. 213, 11899. DOI: https://doi.org/10.1016/j.eswa.2022.118993

[8] Fan, Y., Lam, J.C.K., Li, V.O.K., 2018. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition. In: Kůrková, V., Manolopoulos, Y., Hammer, B., et al. (eds.). Artificial Neural Networks and Machine Learning – ICANN 2018. Springer: Cham, Switzerland. pp. 84–94. DOI: https://doi.org/10.1007/978-3-030-01418-6_9

[9] Elakkiya, R., Natarajan, B., 2021. ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition. Mendeley Data. DOI: https://doi.org/10.17632/kcmpdxky7p.1

[10] Girija, V.R., Sudha, T., Cheriyan, R., 2023. Analysis of Sentiments in Low Resource Languages: Challenges and Solutions. Proceedings of the 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE); 08–11 November 2023; Kerala, India. pp. 1–6. DOI: https://doi.org/10.1109/RASSE60029.2023.10363469

[11] Venkit, P.N., Wilson, S., Cheriyan, R., 2021. IDENTIFICATION OF BIAS AGAINST PEOPLE WITH DISABILITIES IN SENTIMENT ANALYSIS AND TOXICITY DETECTION MODELS. arXiv: Cornell Tech, New York, USA. pp. 1–12. DOI: https://doi.org/10.48550/arXiv.2111.13259

[12] Mgimwa, P.A., Dash, S.R., 2024. Reviving Endangered Languages: Exploring AI Technologies for the Preservation of Tanzania's Hehe Language. In: Mohanty, S.S., Dash, S.R., Parida, S. (eds). Applying AI-Based Tools and Technologies Towards Revitalization of Indigenous and Endangered Languages. Studies in Computational Intelligence. 1148. Springer: Singapore. DOI: https://doi.org/10.1007/978-981-97-1987-7_2

[13] Elakkiya, R., Vijayakumar, P., Kumar, N., 2021. An optimized generative adversarial network-based continuous sign language classification. Expert Systems with Applications. 182, 115276. DOI: https://doi.org/10.1016/j.eswa.2021.115276

[14] Zhou, H., Zhou, W., Zhou, Y., et al., 2020. Spatial-temporal multi-cue network for continuous sign language recognition. Proceedings of the AAAI conference on artificial intelligence. 34(07), 13009–13016.

[15] Jindal, N., Yadav, N., Nirvan, N., et al., 2022. Sign Language Detection using Convolutional Neural Network (CNN). Proceedings of the 2022 IEEE World Conference on Applied Intelligence and Computing; 17–19 June 2022; Sonbhadra, India. pp. 354–360.

[16] Bantupalli, K., Xie, Y., 2018. American Sign Language Recognition using Deep Learning and Computer Vision. Proceedings of the 2018 IEEE International Conference on Big Data (Big Data); 10–13 December 2018; Seattle, WA, USA. pp. 4896–4899.

[17] Ng, R., Zou, E., Ahn, H.S., 2021. Sign Language and Emotion Understanding. Proceedings of the HRI '21 Companion: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction; 8–11 March 2021; Boulder CO, USA. pp. 673–674.

[18] Li, D., Opazo, C. R., Yu, X., et al., 2020. World Level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison. Proceedings of the IEEE Winter Conference on Application of Computer Vision (WACV); 01–05 March, 2020; Snowmass, CO, USA. DOI: https://doi.org/110.1109/WACV45572.2020.9093512

[19] Mariappan, M., Gomathi, V., 2019. Real-Time Recognition of Indian Sign Language. Proceedings of the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS); 21–23 February 2019; Chennai, India. pp. 1–6.

[20] Kamal, S.M., Chen, Y., Li, S., 2017. Technical Approaches to Chinese Sign Language Processing: A Review. IEEE Access. 7, 96926–96935.

[21] Abhishek, K.S., Qubeley, L.C.F., Ho, D., 2016. Glove Based Hand Gesture Recognition Sign Language Translation Using Capacitive Touch Sensor. Proceedings of the 2016 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC); 03–05 August 2016; Hong Kong, China. pp. 334–337.

[22] Xiao, Q., Qin, M., Guo, P., et al., 2019. Multimodal Fusion Based on LSTM and a Couple Conditional Hidden Markov Model for Chinese Sign Language Recognition. IEEE Access. 7, 112258–112268.

[23] Rosero-Montalvo, P.D., Godoy-Trujillo, P., Flores-Bosmediano, E., et al., 2018. Sign Language Recognition Based on Intelligent Glove Using Machine Learning Techniques. Proceedings of the 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM); 15–19 October 2018; Cuenca, Ecuador. pp. 1–5.

[24] Wang, H., Chai, X., Hong, X., et al., 2016. Isolated sign language recognition with Grassmann covariance matrices. ACM Trans. ACM Transactions on Accessible Computing (TACCESS). 8(4), 14.

[25] Liang, Z., Liao, S.-B., Hu, B.-Z., 2018. 3D convolutional neural networks for dynamic sign language recognition. The Computer Journal. 61(11), 1724–1736.

[26] Miah, A.S.M., Hasan, M.A.M., Shin, J., et al., 2023. Multistage spatial attention-based neural network for hand gesture recognition. Computers. 12(1), 13.

[27] Hu, H., Zhao, W., Zhou, W., et al., 2021. SignBERT: Pretraining of hand-model-aware representation for sign language recognition. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 10–17 October 2021; Montreal, QC, Canada. pp. 11067–11076. DOI: https://doi.org/10.1109/ICCV48922.2021.01090

[28] Pugeault, N., Bowden, R., 2011. Spelling it out: Real-time ASL fingerspelling recognition. Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops); 06–13 November 2011; Barcelona, Spain. pp. 1114–1119.

[29] Koller, O., Ney, H., Bowden, R., 2016. Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labeled. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27–30 June 2016; Las Vegas, NV, USA. pp. 3793–3802.

[30] von Agris, U., Knorr, M., Kraiss, K.-F., 2008. The significance of facial features for automatic sign language recognition. Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition; 17–19 September 2008; Amsterdam, Netherlands. pp. 1–6.

[31] Rastgoo, R., Kiani, K., Escalera, S., et al., 2021. Sign Language Production: A Review. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 19–25 June 2021; Nashville, TN, USA. pp. 3446–3456.

[32] Jin, C.M., Omar, Z., Jaward, M.H., 2016. A mobile application of American sign language translation via image processing algorithms. Proceedings of the 2016 IEEE Region 10 Symposium (TENSYMP); 09–11 May 2016; Bali, Indonesia. pp. 104–109.

[33] Al-Barahamtoshy, O.H., Al-Barhamtoshy, H.M., 2017. Arabic text-to-sign (ArTTS) model from automatic SR system. Procedia Computer Science. 117, 304–311.

[34] Tolba, M.F., Elons, A.S., 2013. Recent Developments in Sign Language Recognition Systems. Proceedings of the 2013 8th International Conference on Computer Engineering & Systems (ICCES); 26–28 November 2013; Cairo, Egypt. pp. 36–42.

[35] Kumar, P., Roy, P.P., Dogra, D.P., 2018. Independent Bayesian classifier combination-based sign language recognition using facial expression. Information Sciences. 428, 30–48.

[36] Jain, N., Kumar, S., Kumar, A., et al., 2018. Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters. 115, 101–106.

[37] Barsoum, E., Zhang, C., Ferrer, C.C., et al.,

2016. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. Proceedings of the 18th ACM International Conference on Multimodal Interaction; 12–16 November 2016; Tokyo, Japan. pp. 279–283. DOI: https://doi.org/10.1145/2993148.2993165

[38] Jung, H., Lee, S., Yim, J., et al., 2015. Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 07–13 December 2015; Santiago, Chile. pp. 2983–2991.

[39] Zheng, X., Guo, Y., Huang, H., et al., 2020. A survey of deep facial attribute analysis. International Journal of Computer Vision. 128, 2002–2034.

[40] Savran, A., Gur, R., Verma, R., 2013. Automatic detection of emotion valence on faces using consumer depth cameras. Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops; 02–08 December 2013; Sydney, NSW, Australia. pp. 75–82.

[41] Kelly, D., Delannoy, J.R., McDonald, J., et al., 2009. A framework for continuous multimodal sign language recognition. Proceedings of the 2009 International Conference on Multimodal Interfaces; 2–4 November 2009; Cambridge, Massachusetts, USA. pp. 351–358.

[42] Yang, H.-D., Lee, S.-W., 2011. Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. Proceedings of the 2011 International Conference on Machine Learning and Cybernetics; 10–13 July 2011; Guilin, China. pp. 1726–1731.

[43] Yang, H.-D., Lee, S.-W., 2013. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. Pattern Recognition Letters. 34(16), 2051–2056.

[44] Huang, J., Zhou, W., Zhang, Q., et al., 2018. Video-Based Sign Language Recognition Without Temporal Segmentation. Thirty-Second AAAI Conference on Artificial Intelligence. 32(1). DOI: https://doi.org/10.1609/aaai.v32i1.11903

[45] Aran, O., Burger, T., Caplier, A., et al., 2009. A belief-based sequential fusion approach for fusing manual signs and non-manual signals. Pattern Recognition. 42(5), 812–822.

[46] Oguike, O., Primus, M. Multimodal Sentence Level Sentiment Analysis of Sign Language Using Deep Learning and Language Model. Available from: https://authorea.com/users/824042/articles/1220511-multimodal-sentence-level-sentiment analysis-of-sign-language-using-deep-learning-and-language-models (cited 30 August 2024).

[47] Degadwala, S., Vyas, D., 2024. Survey on Systematic Analysis of Deep Learning Models Compare to Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 10(3), 556–566. DOI: https://doi.org/10.32628/CSEIT24103206

[48] Pandeya, Y.R., Bhattarai, B., Lee, J., 2021. Deep-Learning-Based multimodal emotion classification for music videos. Sensors. 21(14), 4927.

[49] Dashtipour, K., Gogate, M., Cambria, E., et al., 2021. A novel context-aware multimodal framework for Persian sentiment analysis. Neurocomputing. 457, 377–388.