Information Processing and Management 59 (2022) 102929

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks

Alireza Ghorbanali [a], Mohammad Karim Sohrabi [a,*], Farzin Yaghmaee [a,b]

[a] Department of Computer Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran
[b] Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran

ARTICLE INFO

Keywords:
Multimodal sentiment analysis
Transfer learning
Ensemble learning
Deep learning
Dempster-Shafer

ABSTRACT

Huge amounts of multimodal content and comments in a mixture form of text, image, and emoji are continuously shared by users on various social networks. Most of the comments of the users in these networks have emotional aspects, which make the multimodal sentiment analysis (MSA) an important and attractive research topics in this area. In this paper, an ensemble transfer learning method is exploited to propose a hybrid MSA model based on weighted convolutional neural networks. The extended Dempster–Shafer (Yager) theory is also utilized in the proposed method of this paper to fuse the outputs of text and image classifiers to determine the final polarity at the decision level. The pre-trained VGG16 network is firstly used to extract visual features and fine-tune on the MVSA-Multiple and T4SA datasets for image sentiment classification. The Mask-RCNN model is then exploited to determine the objects in the images and convert them to text. The BERT model receives the output of this step along with the textual descriptions of the images for extracting the text features and embedding the words. The output of the BERT model is then imported into a weighted convolutional neural network ensemble (WCNNE). The texts are classified by several weak learners using the AdaBoost that is an ensemble learning technique in which, classifiers are trained sequentially. The combined use of several weak classifiers results in a strong classification. The WCNNE improves the performance and increases the accuracy of the results. As a fusing phase at the decision level, the outputs of the VGG16 and the WCNNE models will be finally merged using the extended Dempster-Shafer theory to obtain the correct sentiment label. The results of the experiments on the MVSA-Multiple and T4SA datasets show that the proposed model is better than the other compared methods and achieved an appropriate accuracy of 0.9348 on MVSA and 0.9689 on the T4SA datasets. Moreover, the proposed model reduces training time due to the use of transfer learning and the proposed AdaBoostCNN achieves better results compared to the single CNN.

1. Introduction

Useful information extracted from sentiments of posts and comments of users in a social network can be used for several purposes (Li, Chen, Zhong, Gong, & Han, 2022; Liu, Li, & Ji, 2021;. The polarities of these sentiments can be divided into positive, negative, and neutral categories. Sentiment analysis can be interpreted as a classification task in which each class represents a sentiment. Sentiment analysis has many applications, including financial forecasts and stock price prediction (Jing, Wu, & Wang, 2021; Li, Shi, Wang, &

* Corresponding author.
E-mail address: Amir_sohraby@aut.ac.ir (M.K. Sohrabi).

https://doi.org/10.1016/j.ipm.2022.102929
Received 15 October 2021; Received in revised form 27 January 2022; Accepted 9 March 2022
Available online 18 March 2022

Zhou, 2021; Xing, Cambria, & Welsch, 2018), politics (Haselmayer & Jenny, 2017), medicine (Chakraborty et al., 2020), and e-tourism (Abbasi-Moud, Vahdat-Nejad, & Sadri, 2021). Sentiment analysis on social media is used for determining public opinion (Behera, Jena, Rath, & Misra, 2021). As social media grows in popularity, most people share their text messages with images or videos, and this visual information is an additional channel for expressing users' feelings. Textual and visual representation of sentiments is a useful structure for better extraction and understanding of sentiments (Kumar, Srinivasan, Cheng, & Zomaya, 2020). Sentiment analysis or opinion mining is one of the research topics in machine learning and natural language processing (NLP). A lot of multimodal comments of several media, such as text and image, are also daily posted on social networks, which creates a huge amount of multimodal data. Analyzing this multimodal data leads to enhance efficiency in understanding comments and opinions (Huang, Zhang, Zhao, Xu, & Li, 2019). Due to the more useful information of the multimodal data compared to the single text or the single image data, it is more likely to recognize the true feelings of users. However, MSA is a challenging task because the information of the sentiments of each modal is different, and it is necessary to properly represent their characteristics.

Several deep learning-based methods have been introduced for MSA in the literature. Deep learning is a machine learning method, the algorithms of which try to learn high-level features from datasets. Automatic feature extraction from raw data, high accuracy in results, and extensive hardware and software support are some of advantages of deep learning. Auto-encoders, deep belief networks (DBNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are some of the most important deep learning models. CNN is one of the most widely used models and is exploited in different fields and applications. The CNNs are utilized in one-dimensional data such as signal and sequence of the words, in two-dimensional data such as image and sound, in three-dimensional data such as high volume videos and images. High accuracy in image recognition and automatic detection of important features without human supervision are two special characteristics of the CNNs. Moreover, since these networks require fewer parameters than the traditional neural network, their memory consumption is reduced and their performance is improved (Liu et al., 2017). CNN has been used for feature extraction in different areas, such as detecting objects. The CNN model more effectively extracts higher-level features using convolutional and max-pooling layers. CNNs are used independently or in combination with the other deep neural networks (DNNs), such as LSTM (long short-term memory), to perform NLP tasks and sentiment analysis (Alayba, Palade, & England, 2018). CNN can also be easily adapted to non-visual data, and thus can be useful for MSA (Sharmin & Chakma, 2021).

Although CNNs are efficient and useful for classification, these networks are computationally heavy, which slows down their training time. Parallel programming methods and the use of processing power of graphics processing units (GPUs), tensor processing units (TPUs) and central processing units (CPUs) not only make the training of large networks possible and faster, but also facilitate simultaneous training of several networks, called ensemble learning. Ensemble learning techniques use a combination of several models to make decisions to increase the model's ability to estimate data output. Ensemble learning algorithms are also able to improve generalization. For example, ensemble learning is used in (Briskilal & Subalalitha, 2022; Dashtipour, Ieracitano, Morabito, Raza, & Hussain, 2021) to reduce the generalization error of domain compatibility and improves accuracy in sentiment classification problems. Using several CNNs leads to high performance (Frazao & Alexandre, 2014) and error rate reduction (Krizhevsky, Sutskever, & Hinton, 2012). Bagging, boosting, and stacking are some of the ensemble learning methods (Zhou, 2021). Ensemble learning are also used in sentiment analysis (Briskilal & Subalalitha, 2022; Sridharan & Komarasamy, 2020).

The huge amount of data needed for training a DNN is an important challenge in this area. In addition to the problems in providing this enormous amount of training data, the very high volume of processing required to train the network with this amount of data is also extremely challenging. Using pre-trained models is an appropriate solution to address this problem. Transfer learning techniques and pre-trained models are used for a variety of NLP tasks, including text classification, sentiment analysis, text production, word embedding, and translation machine. Some of the most popular pre-trained models are BERT, ELMO, and GloVe. The BERT model was introduced by Google engineers in 2018 to perform NLP tasks (Briskilal & Subalalitha, 2022). In general, deep transfer learning examines how DNNs use knowledge from other fields, and a network based on deep transfer learning implies the reuse of a pre-trained network. To solve various challenges and problems, transfer learning techniques are combined with the other algorithms to obtain good results.

A hybrid transfer ensemble learning model is proposed in this paper that uses the pre-trained VGG16 model and fine-tunes it to classify images. It also uses the Mask-RCNN to detect objects in images and convert them to text to improve the accuracy of the polarity detection. The obtained texts along with the image captions are used for word embedding and extracting features in the pre-trained BERT model. The output model is used to extract higher-level features and classify texts imported to a WCNNE. Finally, at the decision level, the Dempster-Shafer (Yager) theory will be utilized to fuse image and text classification outputs to determine the correct polarity of sentiments. Experimental results show that the proposed model of this paper achieves appropriate results. The proposed transfer ensemble learning technique transfers useful samples from the source domain to the target domain to improve classification accuracy. This hybrid transfer ensemble learning model attempts to address some of the important challenges of MSA, including the extraction of useful features from text and images, classification of the extracted features, and fusion of the classification results.

The main features and contributions of this paper are as follow:

1 The pre-trained BERT model is used for word embedding and text features extraction.
2 The Mask R-CNN is utilized to object detection from images and convert them to text.
3 WCNNE is utilized to obtain sentiment polarity from the texts.
4 The pre-trained VGG16 is exploited and fine-tuned for extracting features from images and classifying them.
5 The extended Dempster-Shafer (Yager) theory is used to fuse image and text classification outputs to determine the correct polarity of multimodal sentiments.

The remainder of the paper is organized as follows Section 2. describes the related works on MSA, information fusion, CNN, transfer learning, and ensemble learning method for text sentiment analysis Section 3. explains the details of the methodology of the proposed method. The experimental results are provided and evaluated in Section 4, Section 5 describes the error analysis, and finally, Section 6 concludes the work.

2. Related works and backgrounds

The growth and development of social networks have caused users to share their opinions and ideas on various topics in the form of text, images, audio, and video in these media. The variety of the types of the shared media on social networks has made MSA an important research issue in this area (Li, Zhang, Wang, & Gao, 2021; Zhao et al., 2019). In this section, a brief review of literature in MSA is firstly provided, and then the background of some of related issues, namely CNN, transfer learning, and ensemble learning methods in the context of sentiment analysis will be represented.

2.1. Multimodal sentiment analysis

Social media platform facilitates communication between different communities. Users share their opinions as a combination of text, image, audio, video, or emoji, and thus a huge amount of multimodal data is generated regularly. Single-modal sentiment analysis on text or image may not be complete enough to fully understand the users' emotions. Multimodal sentiment analysis (MSA) is used to better understand emotions, and researchers have paid more attention to it in recent years. An MSA method was proposed in (Huang, Zhang, Zhao, Xu, & Li, 2019) using a combination of CNNs for image classification and LSTM for text classification. In this method, the outputs of the networks were fused in different steps to obtain the correlation between texts and images. A multimodal consistency measure was proposed in (Zhao et al., 2019) for image-text posts. This approach used SentiBank to extract visual features. Pre-trained CNN was used to extract the top-10 tags from images. Two SVM-based models were also used for sentiment prediction. Another approach to classify multimodal sentiments (comments prediction) was presented in (Xi, Xu, Chen, Zhou, & Yang, 2021). By incorporating the utterance-level contextual information and importance of inter-modal utterances, this approach increased the accuracy (Huddar, Sannakki, & Rajpurohit, 2020). Text, video, and audio data were used to sentiment analysis in (Gkoumas, Li, Lioma, Yu, & Song, 2021). The combined methods of decision level and feature level were used in this method to effectively integrate information. The LSTM-trained model was used to extract textual features, the openSMILE model was used to extract audio features, and the 3D-CNN model was used to extract image data features. A two-step method for classifying multimodal sentiments was introduced in (Bawa & Kumar, 2019) using transfer learning and transformer architecture. A new architecture called memory fusion network using LSTM was introduced to analyze sentiments of text and image (Sangeetha & Prabha, 2021). Some studies exploited the combined use of DNNs for MSA. For example, the combined use of CNN and LSTM was used in (Zhang et al., 2018) for MSA.

Video data can also be used as a great resource for analyzing emotions of the users. A lexical knowledge-based extraction approach is developed in (Stappen, Baird, Cambria, & Schuller, 2021) for understanding the content and analysis of emotions from video transcripts, in which, natural language concepts were been extracted using SenticNet, and the extracted features were predicted using support vector machines (SVMs). The users' feelings and interests towards YouTube's movies were used in (Hazarika, Poria, Zimmermann, & Mihalcea, 2021) to extract the features and categorize the emotions. The combination of transcription of the video and the other modalities was utilized in (Majumder, Hazarika, Gelbukh, Cambria, & Poria, 2018) for sentiment classification. An attention-based method was provided in (Marrese-Taylor, Balazs, & Matsuo, 2017) for extracting aspect and topic from videos collected from YouTube. In this method, a small dataset with only 7 videos was used in which high-level concept features were not considered. In (Morency, Mihalcea, & Doshi, 2011), 47 videos were collected and manually transcribed to analyze multimodal sentiments. Multimodal sentiments were analyzed in (Poria, Cambria, & Gelbukh, 2015) on videos using a combination of sound, face state, and text. The text2vec model and CNN network were used to extract the features of the texts in this paper, and The SVM was performed for sentiment analysis. Some related studies in the field of scene segmentation (Chen et al., 2020), movie genre classification (Yadav & Vishwakarma, 2020), video activity recognition (Mliki, Bouhlel, & Hammami, 2020) and movie recommendation (Chen, Yeh, & Ma, 2021) were also carried out in this area.

Reinforcement learning is one of the important categories of machine learning, which has been utilized in various areas of NLP, including sentiment analysis. Environment, agent, reward, state, and a set of possible actions for each state are the main components of reinforcement learning (Duan, Ying, Yuan, Cheng, & Yin, 2021). A reinforcing learning-based approach for analyzing multimodal sentiments (text and sound) was proposed in (Zhang, Li, Zhu, & Zhou, 2019). Deep learning was used in this paper to find clause-level structure in an utterance to model a hierarchical interactive representation for MSA. The use of clause-level information in this paper was similar to (Zhang, Huang, & Zhao, 2018). Using deep learning, Chen et al. performed MSA in (Chen, Wang, Liang, Baltrušaitis, & Zadeh, 2017) on the CMU-MOSI dataset, which is a collection of online videos in which the speaker expresses his or her views on the videos. The attention layer and the input gate controller were trained by reinforcement learning to overcome the noise of modalities. A multimodal reinforcement trading system was introduced in (Chen & Huang, 2021) to improve the performance of the model. Using the development of a robot, the basic interaction skills were successfully learned in (Qureshi, Nakamura, Yoshikawa, & Ishiguro, 2016) through multimodal reinforcement learning.

Multimodal fusion integrates results obtained from multiple data sources to predict the final class value. Early fusion (Gkoumas et al., 2021; El-Sappagh et al., 2021), intermediate fusion (Gkoumas et al., 2021), and late fusion (Xiao, Codevilla, Gurram, Urfalioglu, & López, 2022) are various fusion methods for MSA. Early fusion leads to the production of large input vectors and may be redundant. Intermediate fusion takes place in the middle layers of neural networks. The late fusion aggregates the classification decisions.

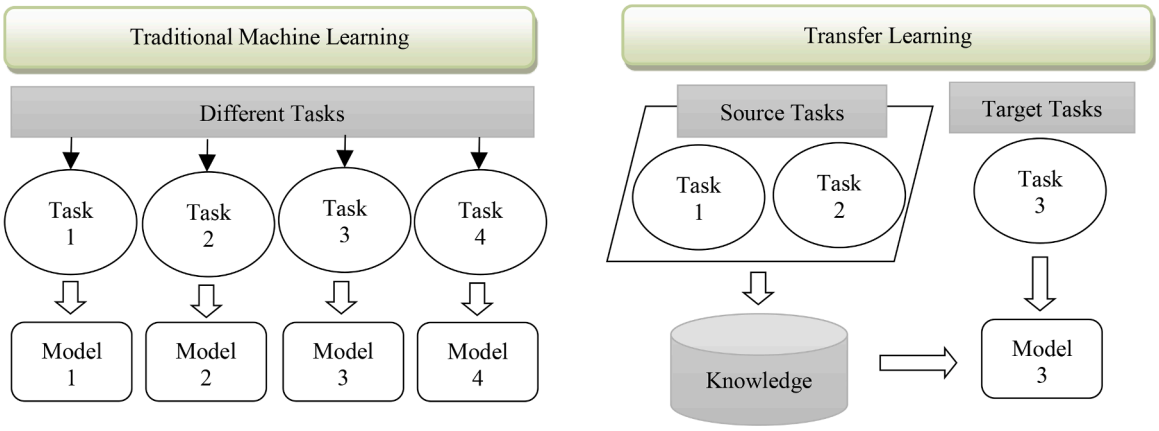

Fig. 1. Comparison of traditional methods and transfer learning techniques.

Multimodal systems often work better than single-modals because of correlations and differences between methods (Huang et al., 2019). A hierarchal attention-iLSTM model was presented in (Huddar, Sannakki, & Rajpurohit, 2021) based on the cognitive brain limbic system for MSA. The hash algorithm was utilized in this method used to improve the retrieve accuracy and speed. Late fusion was used in (Huang et al., 2019) to combine the classifications of MSA.

2.2. Convolutional neural network

CNN or ConvNet is a hierarchical DNN introduced in (Hubel & Wiesel, 1962). The CNN was used in 1998 for the famous MNIST handwritten digit identification project and achieved promising results (LeCun, Bottou, Bengio, & Haffner, 1998). The CNN usually consists of several blocks. The different layers or blocks on the CNN are convolutional layer, non-linear activation function, pooling layer, and fully connected layer. The ConvNet algorithm requires less preprocessing than the other classification algorithms. The ConvNet architecture is similar to the pattern of neuronal connections in the human brain. CNNs are one of the most widely used deep learning architectures for image processing and recognition, but they can also be used for text classification and NLP (Pavel et al., 2021; Sharma, Chaurasia, & Srivastava, 2020). A hybrid architecture using CNN and Bi-GRU networks was proposed in (Cheng et al., 2020) to analyze text emotions. The attention mechanism was used in this architecture to discover more appropriate features. In (Ombabi, Ouarda, & Alimi, 2020), a hybrid method was provided for analyzing emotions with CNN-LSTM networks on Arabic language datasets, in which, local features were extracted using CNN. Two layers of LSTM were also used to maintain text dependencies. A hybrid model of CNN and the genetic algorithm was also used in (Ishaq, Asghar, & Gillani, 2020) to text sentiment analysis. Other hybrid methods in the field of sentiment analysis can be considered in (Behera et al., 2021; Meškelė & Frasincar, 2020; Pandey, Rajpoot, & Saraswat, 2017). Image sentiment analysis was also accomplished on the social network dataset in (Salunke & Panicker, 2021) and achieved a proper accuracy. CNN networks are great for extracting local features from the text. To classify texts using CNN networks, text data can be entered into the CNN network using pre-trained word vectors. To do this, each word of the text is considered as a word vector. The convolution layers are used to extract the implicit features from the inputs or intermediate features map. The length of each filter is the constant value of K and its height is the hyper-parameter h. Given a filter $W \in \mathbb{R}^{h*k}$, a feature C_i is generated from the window of words $[V^i : V^{i+h-1}]$, which refers to the sequence $[X^i \oplus X^{i+1} \oplus X^{i+2} \oplus \ldots \oplus X^{i+h-1}]$ as shown in Eq. (1).

$$C_i = g(W.[V^i : V^{i+h-1}] + b \tag{1}$$

In this Equation, $b \in \mathbb{R}$ is a bias term and g is a non-linear function. ReLu is used in this paper as the non-linear function for the convolution layer. The filter W is applied to each possible window of words in the sequence $[v^{1:h}, v^{2:h+1}, \ldots, v^{n-h+1}]$. The result of applying this filter is the feature map of Eq. (2), in which, $c \in \mathbb{R}^{n-h+1}$.

$$c = [c^1, c^2, \ldots, c^{n-h+1}] \tag{2}$$

This process can be repeated for different filters by different heights (Kim, 2014). Using fully connected networks these extracted features can be used. The proposed method of this paper combines several CNNs to extract high-level features and classify texts.

2.3. Transfer learning

Transfer learning transfers the parameters of a trained neural network over a dataset and reuses them for another problem with another dataset and purpose, and thus transfers knowledge from a source domain to a destination domain to build a proper classification model (Agarwal, Sondhi, Chopra, & Singh, 2021; Smetanin & Komarov, 2021,) Fig. 1. shows the difference between traditional machine learning methods and transfers learning techniques.

The combined use of transfer learning and the other algorithms was exploited to solve some real-world problems. In DNNs, their

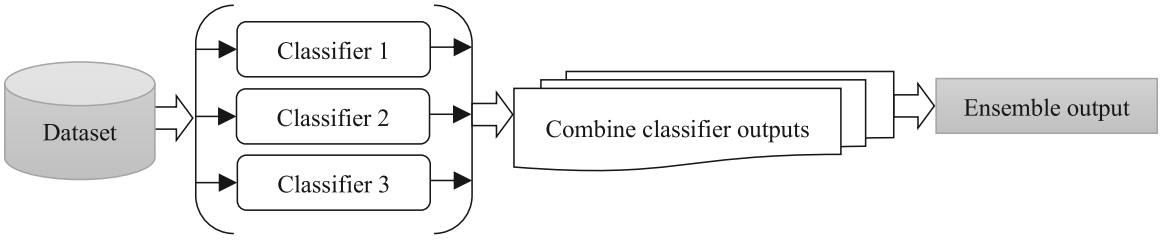

Fig. 2. Ensemble classification framework.

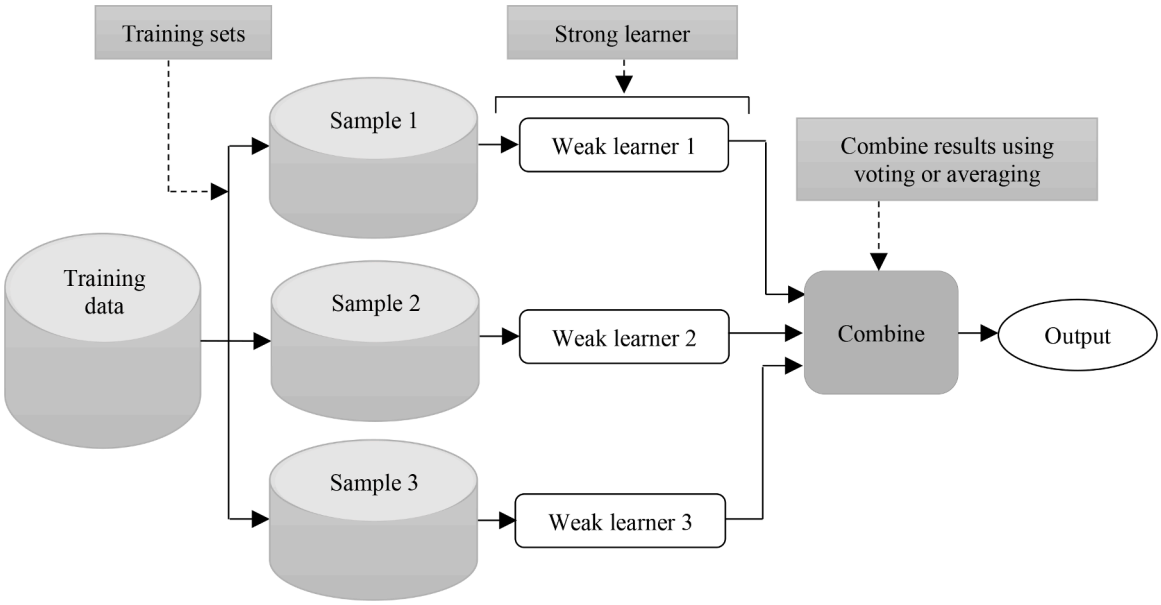

Fig. 3. Bagging method.

first layers learn general features that are not specific to a particular dataset or task, and thus can be used for many datasets and other tasks (Rezende et al., 2018). To improve the power of their deep network, high-level image features were extracted in (Devulapalli, Potti, Krishnan, & Khan, 2021) using a pre-trained Google net model. The sentiments of the images were analyzed in (El-Sappagh et al., 2021)using a transfer learning strategy. The pre-trained VGG16 model was used to extract visual features and the GloVe pre-trained model was used to create features in (Behera et al., 2021) for MSA. In various studies, the pre-trained BERT model has been used for text sentiment analysis (Wang, Lu, Chow, & Zhu, 2020). The results of the research have shown that the use of transfer learning techniques can improve performance in various fields (Zhuang et al., 2020).

2.4. Ensemble learning methods

Neural networks have many applications including knowledge representation (Li et al., 2018), modeling (Belhaj et al., 2021), prediction (Jing et al., 2021), and automated design. All of these applications use an integrated neural network structure. In this integrated structure, neural networks have a single architecture, in which, only one neural network is used to perform tasks. The scalability is one of the main challenges of neural networks (Gepperth & Karaoguz, 2016). Using a combined learning approach instead of using a single neural networks can be suggested to solve the complex problems. A combination of several learning models, called ensemble learning, is used in this approach, which increases the estimation power of the final model. In ensemble learning algorithms, a sample is classified by several different classifiers, the results of the classifications are combined in different ways, and the final result is determined for that particular sample, which usually leads to improve accuracy and performance. In recent years, ensemble learning has attracted a lot of attention in the fields of artificial intelligence, machine learning, neural networks, pattern recognition, and data mining. Several base learners or weak learners are trained to solve a problem, and then will be combined to get better results. When weak models are properly combined, they can create more accurate models or strong learners Fig. 2. shows an ensemble network with 3 classifiers. A typical ensemble classification model consists of two steps, namely generating initial results using several base learners, combining the results obtained by the base learners using methods such as voting or averaging. The use of deep ensemble networks usually leads to more accurate result (Bargshady et al., 2020; Onan, Korukoğlu, & Bulut, 2017).

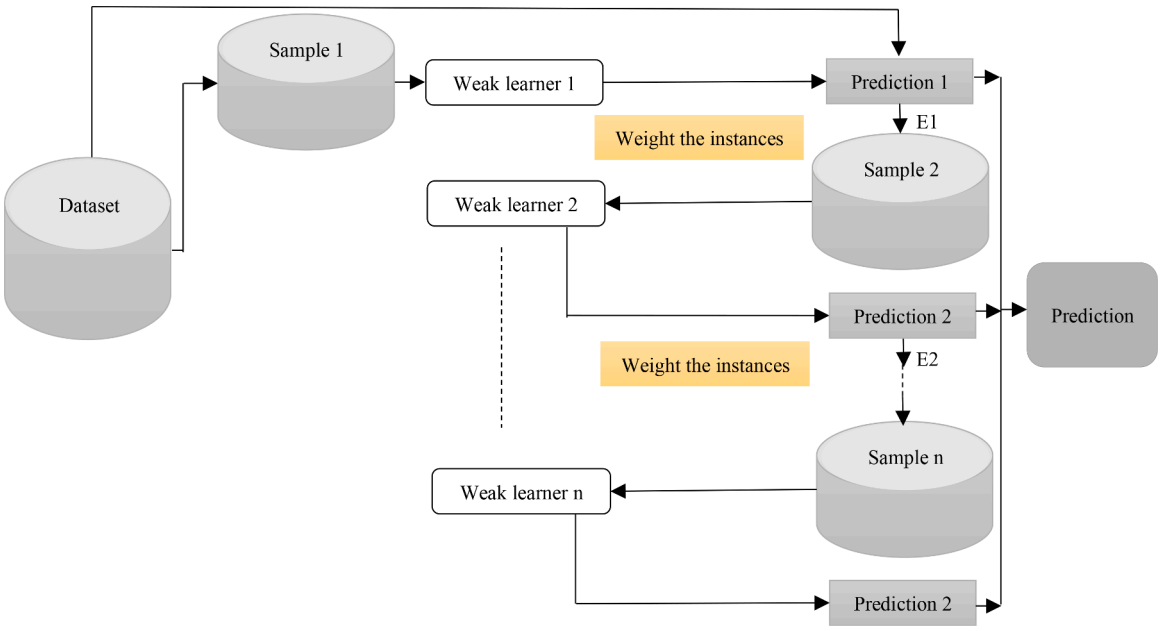

Fig. 4. Boosting method.

Popular ensemble learning techniques include bagging, boosting, and voting. The bagging or bootstrap aggregating method is the simplest and the most effective way to do the learning process of a group of heterogeneous training sets. The Bagging method produces some subsets by random sampling of the training set, and uses these subsets, which may contain duplicate data, to independently learn the basic models in parallel (Dong, Yu, Cao, Shi, & Ma, 2020). The bagging method reduces the variance (Rao, Huang, Feng, & Cong, 2018) Fig. 3. shows the bagging method.

Boosting is one of the most widely used ensemble learning methods that can significantly improve the performance of learning algorithms (Dhamayanthi & Lavanya, 2021), and generally has two main steps: 1) using several weak learners to classify training data, 2) combining the results obtained by the weak learners using a single classifier. The purpose of this method is to do a complex task by combining several weak classifiers. In this method, the basic models are trained sequentially and independently. Two important boosting methods are AdaBoost, and gradient boosting algorithms. AdaBoost is a simple and powerful method that updates the weights attached to each training sample and has been used in several areas including text classification (Onan, Korukoğlu, & Bulut, 2016). In general, weights are assigned to each training set in boosting, and k classifiers are trained repeatedly. After training classifier M_i, the weights are updated to the next classifier, and M_{i+1} learns more about training samples, incorrectly classified by M_i. The final boosting classifier M^* combines the votes of classifiers. The AdaBoost was used in some transfer learning, such as TrAdaBoost, TransferBoost, and TrBagging Fig. 4. shows the boosting method. In (Kazmaier & van Vuuren, 2022), using several heterogeneous classifications, the performance sentiment analysis was improved by 5.53% compared to the single classifier. An ensemble architecture of CNN for image classification was introduced in (Chen et al., 2019), which considered a weight for each classifier according to the produced results. The ensemble CNNs have also been used in various other fields, such as object detection, estimation (Kawana, Ukita, Huang, & Yang, 2018), sentiment analysis (Chatterjee, Mukhopadhyay, Goswami, & Panigrahi, 2021), and stock forecasting. An ensemble model of SVM, CNN, and MLP networks for text sentiment analysis on Persian language datasets was proposed in (Dashtipour et al., 2021).

2.5. *Combining ensemble learning and transfer learning approaches*

Several studies (Huang et al., 2019; Kumar et al., 2020) have used independent classification models separately to extract text and image features to classify them and finally have utilized common fusion methods to combine the results obtained from classification. Some other studies (Huang et al., 2019; Zhao et al., 2019) have used transfer learning methods to extract text and image features, and fine-tuned them for their data. Some studies has exploited ensemble learning methods for MSA (Huddar & Sannakki, 2018; Poria, Peng, Hussain, Howard, & Cambria, 2017). In the proposed method of this paper, a hybrid transfer learning model is presented based on WCNNE along with a probabilistic method for determining the polarity of multimodal data. To improve accuracy, the proposed model uses an additional input. Using the Mask-RCNN, the objects are detected in the image and converted into text for better analysis of the sentiments. In general, the proposed method of this paper extracts appropriate features from the text and image and obtains better results using ensemble classification. Transfer learning and ensemble learning have many advantages. This research is benefited from the advantages and features of the both techniques to improve the performance, reduce the training time and save the resources. For example, the most important features of transfer learning are resource savings, improving efficiency in training new models, and no need for large amounts of data. With ensemble learning, better predictions are made and the model performs better than single

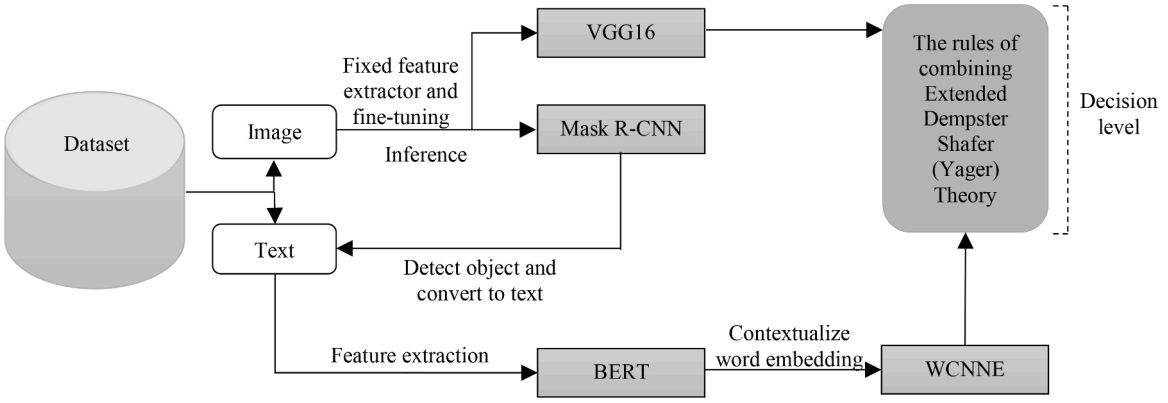

Fig. 5. The proposed architecture for MSA.

Table 1
An example of raw and cleaned tweets in T4SA dataset (Vadicamo et al., 2017), before and after pre-processing.

id	Text	Cleaned text
758014713804587008	RT @polarcomic: And surprise! the #RegularShow...	RT polarcomic And surprise the RegularShow ...
758014717990428672	RT @SweetBabyBellB: My unproblematic fav who ...	RT SweetBabyBellB My unproblematic fav who...
758014646716665857	RT @WhyLarryIsReal: I mean we know harry isn't...	RT WhyLarryIsReal I mean we know harry isnt...
758014655071526912	RT @Eastbay: She's ready, resilient, and on ou...	RT Eastbay Shes ready resilient and on...
758014642526429184	RT @SheeeRatchet: find someone who loves you a...	RT SheeeRatchet find someone who loves you a...

classifier. Prediction is improved, and variance and bias are reduced using bagging and boosting methods, respectively.

3. The proposed method

In this study, the combined use of transfer learning and ensemble learning techniques is exploited to increase the accuracy and improve the performance of MSA. Fig. 5 shows the architecture of the proposed method of this paper. As shown in Fig. 5, two different methods of fine-tuning and feature extraction are used on the VGG16 model to classify the image data. Furthermore, using the inference on the Mask-RCNN model, the objects in the image are detected and converted to text to increase the accuracy of sentiment classification. Extracted text from the images and text captions of the images are given to the pre-trained BERT model to discover features and embedded words. The conceptual features of the words are then entered into the proposed WCNNE to classify the texts. To determine the final polarity, the outputs of the fine-tuned VGG16 model and WCNNE are fused using extended Dempster-Shafer (Yager) theory at the decision level.

3.1. Data preparation

The dataset must be prepared before entering the neural networks. The dataset consists of text and image parts, each of which are prepared separately.

- **Cleaning the text data**

The datasets used in this study contain some raw data with acronyms, spelling mistakes, and unwanted symbols. In the pre-processing phase, unwanted symbols such as !, @, &, #, and numbers (2, 4, .) are removed Table 1. shows an example of raw and cleaned tweets in T4SA dataset. Tokenize operations are used to separate words. The BERT tokenizer is utilized to break the raw text down into tokens.

- **Feature extraction for image and text**

In this study, feature extraction is performed automatically using pre-trained models such as VGG16 and BERT. Extraction of appropriate and high-level features leads to improved accuracy and reduced processing time for large inputs. Therefore, the use of pre-trained models will be very effective.

- **Embedding layer**

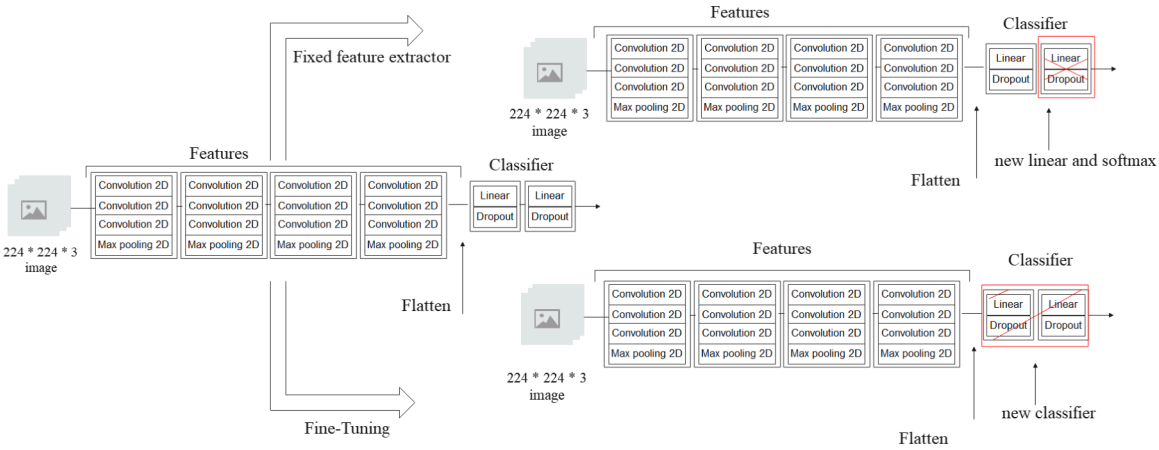

Fig. 6. Feature extraction and fine-tuning in the VGG16 model.

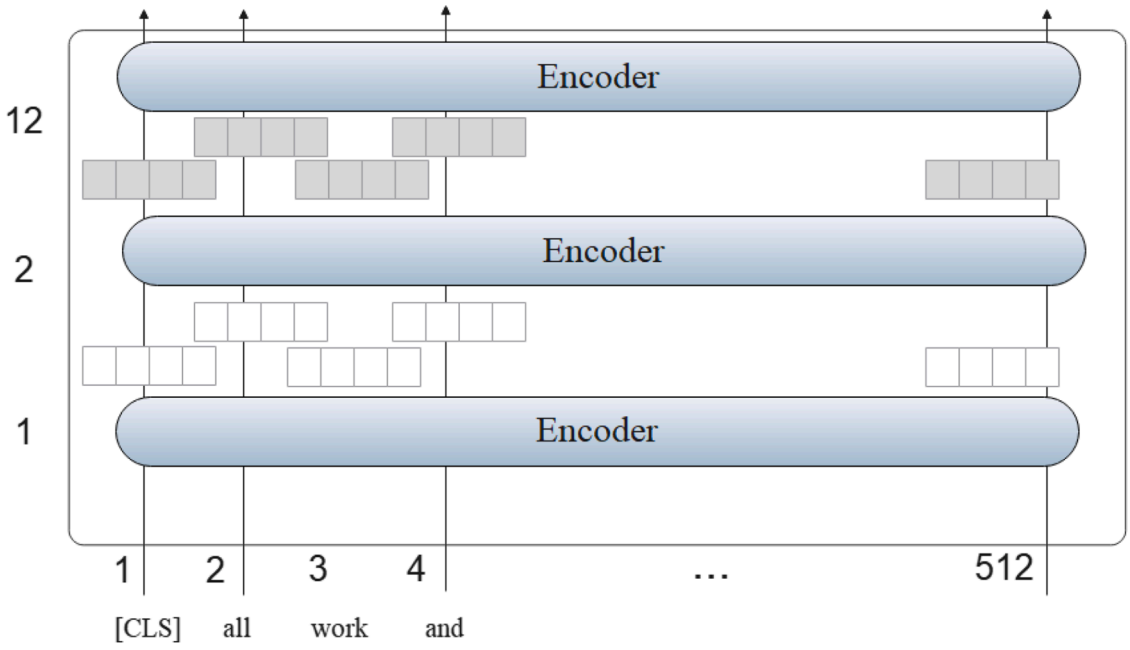

Fig. 7. BERT model architecture.

The BERT model is utilized in the proposed method of this paper to embed words and extract features. Embedding layers convert input texts into vectors.

3.2. Classification of images using transfer learning with fine-tuning the VGG16 model

VGG16 is a CNN model proposed in (Simonyan & Zisserman, 2014), which consists of 13 Convolution and 3 fully connected layers, and is connected to the pooling layer after each step (Qu, Mei, Liu, & Zhou, 2020). The convolution layers use 3×3 kernels with stride 1 and padding 1 to ensure that each activation map retains the same spatial dimensions of the previous layer. A ReLU activation is performed just after each convolution at the end of max pooling. This is done at the end of each block to reduce the spatial dimension. The Max pooling layer with 2×2 kernel and stride 2 is used without padding to ensure that each activation map halves the spatial dimension of the previous layer. Fully connected layers with 4096 ReLU enabled units are used before 1000 fully connected softmax layers (Rezende et al., 2018) Fig. 6. shows the feature extraction and fine-tuning in the VGG16 model.

The Keras library is used to extract features and fine-tunes this model. In the feature extraction phase, the trained network parameters are frozen and the last layer of the fully connected part is removed, the logarithm of the softmax linear classifier is added, and the last layer parameters are trained with the dataset. In the fine-tuning phase, the trained parameters of the feature section are frozen,

Table 2
Parameter settings for the BERT model.

Parameter	Value
Batch size	32
Learning rate	2e − 5
Optimization	adam
MAX_LENGTH(Sentence)	128

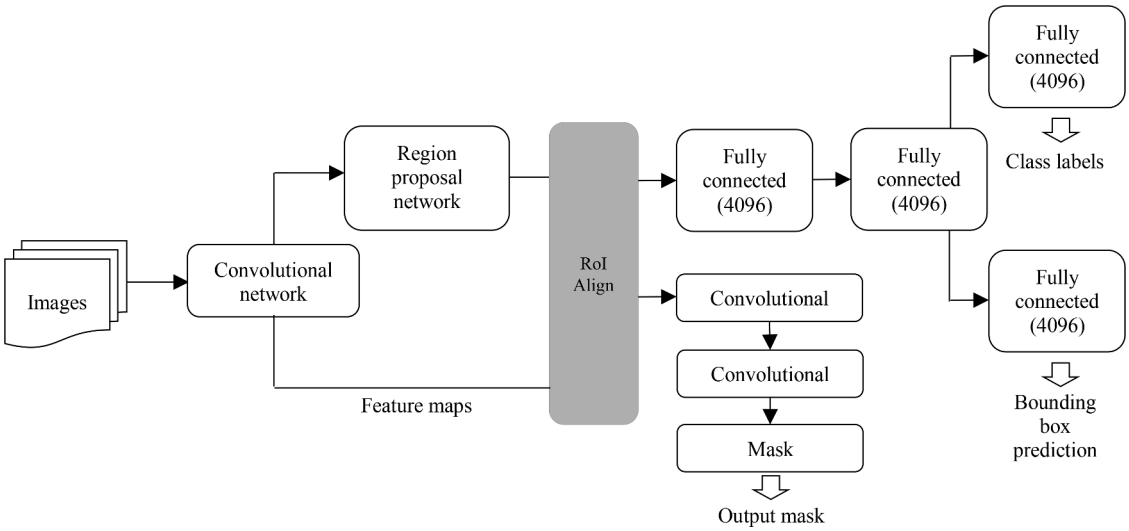

Fig. 8. Architecture of the Mask-RCNN.

a new classifier is added to the fully connected section, and the model is trained on the dataset.

Since the accuracy of the fine-tuning of the VGG16 model on the MVSA and T4SA datasets (with respectively 84% and 85%) is better than accuracy of its feature extraction on these datasets (with respectively 81% and 83%), the fine-tuning method is used as the best classifier. For example, the fine-tuning is performed separately and with different configuration on the T4SA. In this configuration, flatten, dropout, dense, dropout and dense layers are respectively used after the last layer of VGG16.

3.3. Word embedding with BERT

There are different converting methods of word including bag of words, such as count vectoriser and TF-IDF, embedding words as vectors, such as Word2Vec, GloVe, and Doc2Vec, and transformer-based models such as universal sentence encoder and BERT. The BERT model is used for text feature extracting and word embedding using the Keras library in Python Fig. 7. shows the architecture of the BERT model, which can train a text, and is trained on more than 3 million words from the English Wikipedia and the BookCorpus dataset. The BERT model can be fine-tuned by adding an extra output layer for classification issues (Briskilal & Subalalitha, 2022). The BERT model encodes the syntactic and semantic information of the embedding that is necessary to perform a large number of tasks. The BERT model makes the benefit of transformer as an attention mechanism that learns contextual relations among words or sub-words in a text. Pre-training on a variety of data, context-sensitivity, and publishing as an open-source model are some of the main advantages of the BERT.

Since the BERT model is trained on a large amount of data, it can be applied to other datasets and can achieve proper results. Moreover, the BERT model considers different vectors for each word and thus is a context-sensitive model. Since the BERT model returns different embedding for each word, it distinguishes between different uses and meanings of a word. Hence, more information will be available and the performance will increase. Most of the other models, such as Word2Vec and GloVe, create context-independent embedding, and combine all different meanings of a word into one vector. The BERT model also considers the position of each word in the sentence. The BERT is an unsupervised, deep and bidirectional model that has better performance than its previous methods. Finally, since the BERT model is open source it can be customized to extract high-level language features from the text database, and can be also fine-tuned for classification and recognition.

The raw text data is the input of the BERT model. Each word input is embedded in a token using the raw BERT embed. The input embedding of BERT model contains of 3 sections: 1-token embedding, 2-position embedding, 3-segment embedding, and all three have the same dimensions. Also the input embedding is the summation of these three embedding. The first token of any sequence is always a special token [CLS], and special character [SEP] show end of the sentence. In this study, the BERT model uncased_L-12_H-768_A-12 is used by the primary parameters of the sequence encoder Table 2. shows the parameters used in the BERT model. The hyper-parameters

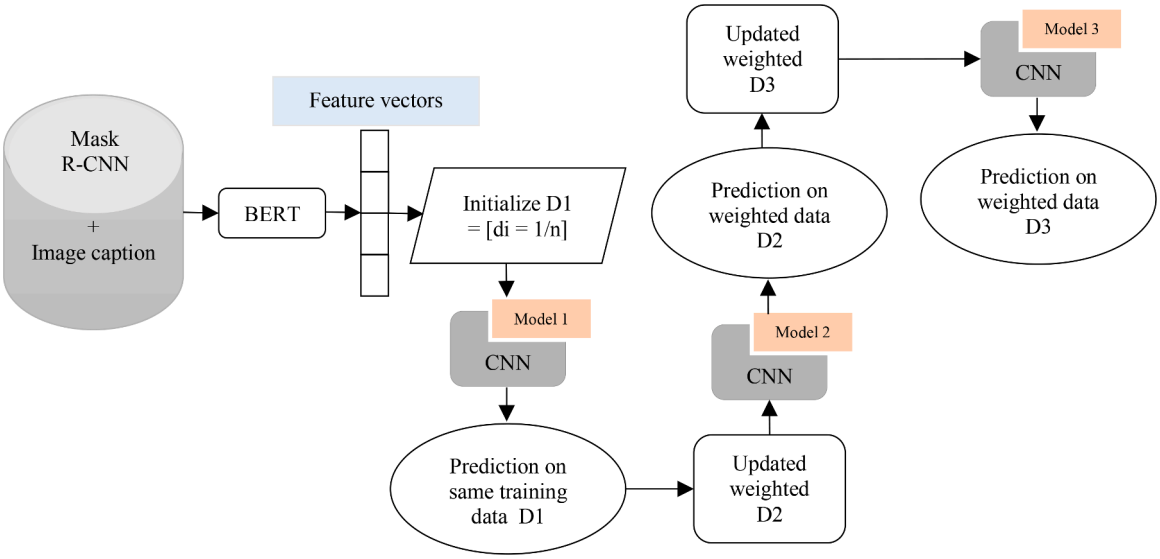

Fig. 9. The proposed transfer ensemble learning network based on BERT-AdaBoost-CNN for text classification.

can be fine-tuned by values 8, 16, 32, 64, and 128 for the batch size, 3e-4, 1e-4, 5e-5, 3e-5, and 2e-5 for the learning rates, and 2, 3, and 4 for the number of epochs.

3.4. The use of Mask-RCNN model architecture and structure

Mask-RCNN, used for semantic segmentation, segment object localization, and object segmentation (Ahmed, Gulliver, & alZahir, 2020),consists of two steps, namely region proposal network (RPN), which scans the initial feature map and generates the proposed region or area (RoI) suggestions, and RoI pooling, applied to each RoI to sample the feature map using the nearest neighbor approach. Since the RoI pooling can lead to a mismatch between the RoI and the extracted features, RoI-align is applied to each RoI to create more accurate RoIs. Both Mask-RCNN steps are connected to the backbone structure, which is a DNN used to map early features. The backbone network can be a CNN trained on an image dataset (Zhuang et al., 2020) Fig. 8. shows the architecture of the Mask-RCNN model. The Mask-RCNN model is used in inference mode, and using this trained model, the objects are detected in the images, are converted into the text, and the obtained text is added to the text captions of the images.

3.5. The proposed WCNNE using AdaBoost

The results of experiments and research have shown that the use of ensemble CNNs extracts much higher quality features than traditional CNNs (Dong et al., 2020). In this study, a WCNNE is used for text classification Fig. 9. shows the proposed transfer ensemble learning network for text classification. The pre-trained BERT model is used in this network to extract features and word embedding, and its output is then used for the final classification of the texts by the WCNNE. In the proposed ensemble CNN, the AdaBoost technique is used to weight training samples. AdaBoost is one of the popular boosting methods. Assume the dataset D as a set of d labeled data, (x_1, y_1), (x_2, y_2), …, (x_d, y_d), where y_i is the label of x_i. Using AdaBoost, the same weight $1/d$ is firstly assigned to each training sample. For the ensemble that requires k rounds via the rest of the algorithm, k classifiers should be utilized. In step i, a subset D_i of D is considered as a training set. The chance of selecting every set is based on its weight. A learner or classification model C_i is originated from the D_i. The weights of the training samples are adjusted according to their classification.

If the training sample is not classified correctly, its weight will increase, and else its weight will not increase. These weights are used to generate the training samples of the next step. The main idea of this method is to focus the classifier on training samples not classified correctly. To calculate the error rate of the C_i model, the sum of the weights of all training samples in D_i incorrectly classified by C_i is calculated. Eq. (3). shows the error rate of C_i, in which, $error(X_j)$ is the classification error of the training sample X_j.

$$error(C_i) = \sum_{1 \le j \le d} w_j error(X_j) \tag{3}$$

If sample X_j is not classified correctly, the value of $error(X_j)$ will be 1 and else will be 0. If the overall performance of the C_i classifier is so weak that the error exceeds 0.5, it will be removed and a new D_i training set is selected for creating a new C_i model. The C_i error rate affects the update of the weights of training samples. If a training sample is correctly classified in step i, its weight is updated as shown in Eq. (4).

Table 3
Configuration of the CNN used for MVSA (text) dataset.

Layers	Configuration
Conv1D	128, 5, Relu
max-pooling	5
Conv1D	128, 5, Relu
max-pooling	5
Dropout	0.2
Flatten	
fully connected	softmax

Table 4
Configuration of the CNN used for T4SA (text) dataset.

Layers	Configuration
Conv1D	128, 5, Relu
max-pooling	5
Conv1D	128, 5, Relu
max-pooling	5
Conv1D	128, 5, Relu
max-pooling	5
Dropout	0.5
Flatten	
Dropout	0.5
fully connected	softmax

Table 5
Hyper-parameters used in experiments.

Hyper-parameter	Value
batch_size	(32, 128)
Epochs/estimators	(10, 15, 25, 50)
learning_rate	0.01
momentum	0.9
dropout	(0.2, 0.5)
Word dimension	(128, 300)

$$w_i = w_i \frac{error(C_i)}{1 - error(C_i)} \tag{4}$$

Therefore, the weights of incorrectly classified samples will increase and the weights of correctly classified samples will decrease as mentioned earlier. The AdaBoost technique and Eq. (4) are used to calculate the weights of each training set in the proposed method of this paper. In this method, 3 WCNNE are used. CNNs with different parameters and configurations are used for each of the datasets Tables 3. and 4 show the CNN configurations for each of the datasets Table 5. shows the hyper-parameters used in the experiments.

3.6. The Dempster–Shafer theory

Evidence theory was first emerged in 1967 with the publication of Dempster's theory on possible upper and lower bounds (Dempster, 1968) . In 1976, Shafer completed this theory and extended it to analyze non-transparent and incomplete information (Shafer, 1976). Dempster–Shafer theory is based on the values of several evidences (classifiers) of a subject and a combination of the choices of these evidences. Traditional probability theory deal with only one number as the probability, but evidence theory deal with a set of probabilities that may be aligned or may be in conflict. The main advantage of evidence theory is that this method is designed to summarize evidences at different levels of belief values and does not require additional information. This estimation of evidences of subjects may be highly variable. For example, two evidences may be perfectly aligned, have little in common, or may be conflict. Evidence theory is able to analyze inadequate and inaccurate information (Shafer, 1976).

3.6.1. Basic functions in evidence theory

There are 3 important functions in Dempster–Shafer theory, which are the fundamental function of probability mass (m), the belief function (Bel) and the plausibility function (Pl). The mass function is the belief mapping of an evidence to the existence of state A, which is defined by a number between 0 and 1. The other functions of evidence theory are calculated according to the definition of the mass function. The plausibility and the belief functions are respectively the upper and lower bounds of the probability of occurrence of an event, defined based on the mass function. The actual probability of occurrence of an event A, denoted as $p(A)$, is the value between

Table 6
Details of the MVSA and T4SA datasets.

Dataset	#Positive	#Negative	#Neutral	Data type
MVSA (manually annotated)	1398	724	470	Text + image
T4SA	528,203	335,910	849,428	
T4SA (tweets)	371,341	179,050	629,566	
B-T4SA (image subset of T4SA)	156,862	156,862	156,862	

the values of the plausibility and belief of the event (Shafer, 1976). In evidence theory, the conflict of evidences may lead to a completely erroneous estimate. Several methods have been introduced to address this problem, one of the most efficient of which is Yager introduced and formulated in (Yager, 1986). In this theory, the possibility of conflict of evidences is properly considered. The new mass probability level function, denoted as q, is defined in the Yager method instead of the base mass function, the value of which is greater than 0, i.e. $q(\varnothing) \geq 0$. In real applications, the performance of classifications may have many errors. The detection percentage of the evidences is reduced using an importance factor and made more realistic. The importance factor is the reliability degree of an evidence and explains its weight among the other evidences. If $O_i(A)$ is the estimation of i^{th} evidence of the event of state A and α_i is the weight of i^{th} evidence, the new value of the mass function is defined according to Eq. (5).

$$m_i(A) = \alpha_i O_i(A) \tag{5}$$

The inflectional combination (logical conjunction) of the probability level of different events of evidence is calculated as Eq. (6).

$$q(A) = \sum_{\Delta A_i = A} m_1(A_1) m_2(A_2) \ldots . m_i(A_i) \tag{6}$$

Eq. (7) is used for combining evidence estimations in the Yager method.

$$M(A) = \frac{q(A)}{1 - q(\varnothing)} \tag{7}$$

Yager possible errors and conflict of evidences are shown as Θ. The important factor of Θ is defined as Eq. (8).

$$\Theta_i = 1 - \alpha_i \tag{8}$$

Using Θ, Eq. (6) can be rewritten as Eq. (9).

$$q(A) = \sum_{\Delta A_i = A} m_1(A_1) m_2(A_2) \ldots . m_i(A_i) + \Theta_i m_i \tag{9}$$

In this study, the extended Dempster–Shafer (Yager) rules are exploited to fuse the results of text and image classifiers. The accuracies of the classifiers in the training phase are utilized to adjust α and Θ parameters for them. The use of extended Dempster–Shafer (Yager) needs evidence and weights. The method of calculating the weights is given in Eq. (8). The evidences are the confusion matrices derived from the classifiers. By having weight parameters and evidence, the extended Dempster–Shafer (Yager) can use Eqs. (5)–(9) to fuse the results.

4. Implementation details and experiments

In this section, the results of empirical experiments of the implementation of the proposed method and their evaluations are provided.

4.1. The datasets

Two social text and image datasets are used to evaluate the proposed method of this paper, namely MVSA[87] and T4SA (Vadicamo et al., 2017) Table 6. shows the information of the MVSA and the T4SA datasets for tweet sentiment analysis.

4.1.1. The MVSA dataset

The MVSA (Niu, Zhu, Pang, & In, 2016) is one of the datasets used for evaluation. All of the visual-textual content pairs in the MVSA have been gathered from Twitter. The vocabulary consists of 10 classes masking almost all the fillings of humans. A few emotional words, such as happy and sad, regularly seem within the tweets. The MVSA contains 19600 pairs of images and text tagged by 3 annotators. In this dataset, the real feeling is calculated separately for each case by taking the majority of votes from 3 sentiments (positive, negative, and neutral).

4.1.2. The T4SA dataset

The T4SA is the second dataset of this paper. The total number of tweets in the T4SA dataset is about 3.4 million, corresponding to about 4 million images. This dataset has been collected from Twitter for about 6 months (Vadicamo et al., 2017). Each tweet (text and associated images) has been labeled according to the sentiment polarity of the text (negative = 0, neutral = 1, positive = 2), obtaining a

Table 7
The performance of the proposed method for classifying text and image sentiments on the MVSA dataset.

Approach	Accuracy	F1	Recall	precision
Fine-tuning VGG16	0.84	0.78	0.77	0.80
WCNNE and sentence embedding with BERT	0.9542	0.9544	0.9505	0.9585
Extended Dempster-Shafer (Yager)	0.9348	0.9366	0.9391	0.9343

Table 8
The performance of the proposed method for classifying text and image sentiments on the T4SA dataset.

Approach	Accuracy	F1	Recall	precision
Fine-tuning VGG16	0.8582	0.77	0.79	0.76
WCNNE and sentence embedding with BERT	0.9456	0.9457	0.9414	0.9502
Extended Dempster-Shafer (Yager)	0.9689	0.9647	0.9791	0.9509

Table 9
Values $i\alpha$ and Θi for the Dempster Shafer, Yager information composition rules.

Dataset	Approach	Θ_i	α_i
MVSA	Fine-tuning VGG16	0.16	0.84
	WCNNE and sentence embedding with BERT	0.05	0.95
T4SA	Fine-tuning VGG16	0.15	0.85
	WCNNE and sentence embedding with BERT	0.06	0.94

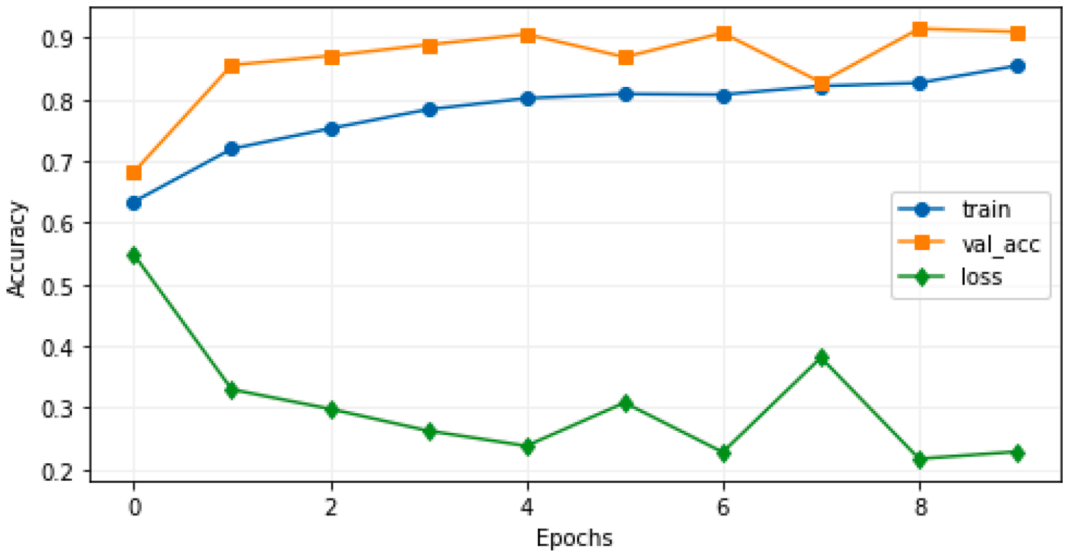

Fig. 10. Accuracy and loss obtained on B-T4SA dataset by fine-tuned VGG16 (Image).

labeled set of tweets and images divided into 3 categories. Corrupted and near-duplicate images have been removed and balanced subset of images, called B-T4SA, has been selected, used to train the visual classifiers.

4.2. Results

Accuracy, precision, F1-score, and recall, respectively calculated using Eqs. (10)–(13), are 4 criteria used to evaluate the efficiency of classification models. Since this classification problem has 3 classes (positive, negative and neutral), the total values of TP (true positive), TN (true negative), FP (false positive), FN (false negative) are calculated and used in these equations.

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

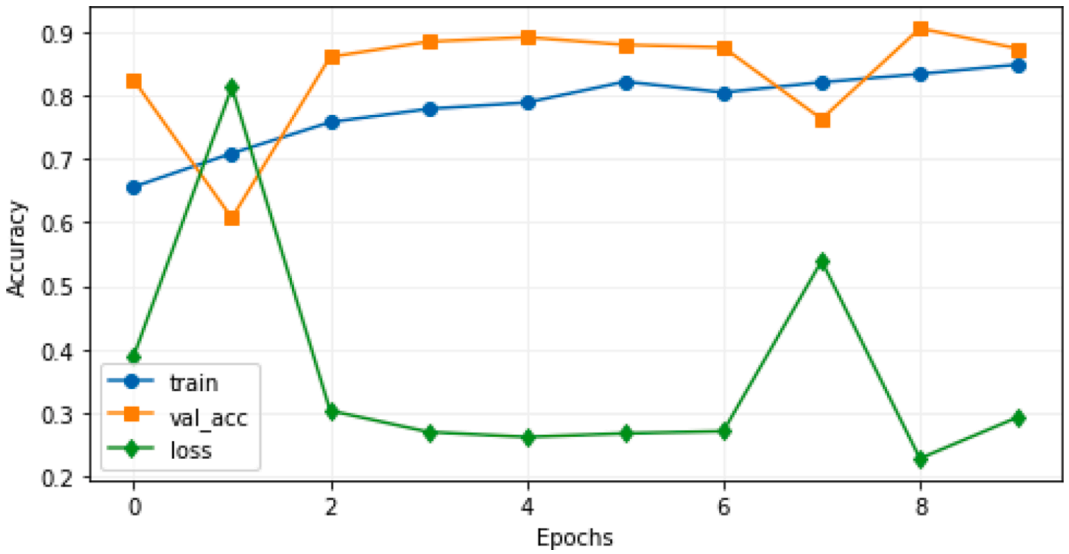

Fig. 11. Accuracy and loss obtained on MVSA dataset by fine-tuned VGG16 (Image).

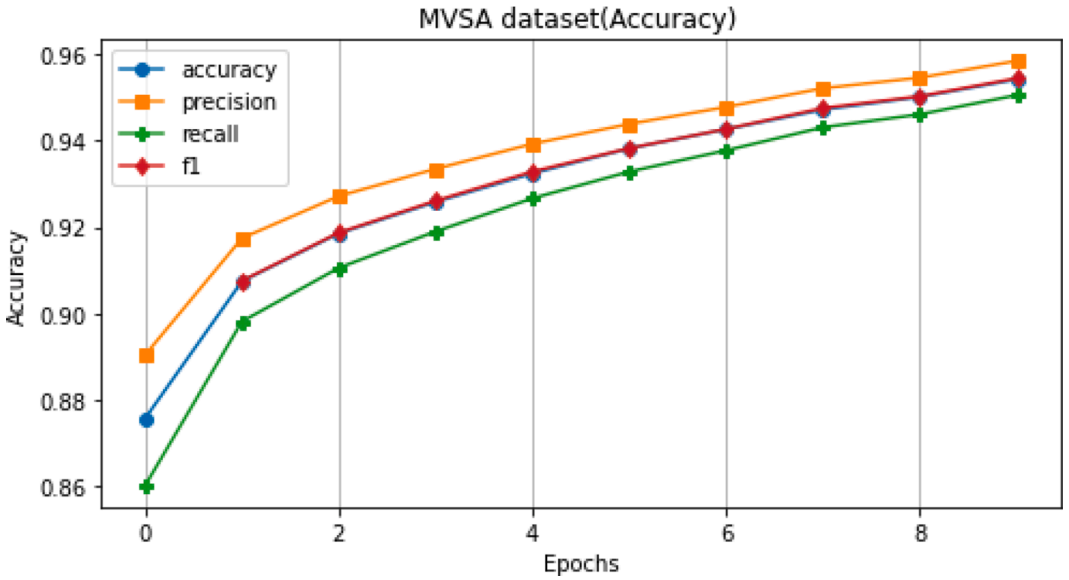

Fig. 12. Accuracy, precision, recall, and F1 on the MVSA dataset (Text).

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

$$F1 - score = \frac{2\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{12}$$

$$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

Tables 7 and 8 show the results of the implementation of the proposed method on the datasets. As shown in Table 7, the use of the combination of results using the Dempster-Shafer, Yager fusion rules has resulted in achieving an appropriate accuracy of 0.93 on the MVSA dataset Table 8. shows the results obtained on the T4SA dataset Table 9. shows the values of the parameters $i\alpha$ and Θi of the Dempster Shafer, Yager data fusion rules according to the accuracy of text and image classification models.

In the Figs. 10 and 11 shows the accuracy of the image classification, and Figs. 12 and 13 shows the accuracy of text classification on MVSA and T4SA datasets. The models are trained in 10 epochs using 70% of the samples as the training set, and 30% of samples are

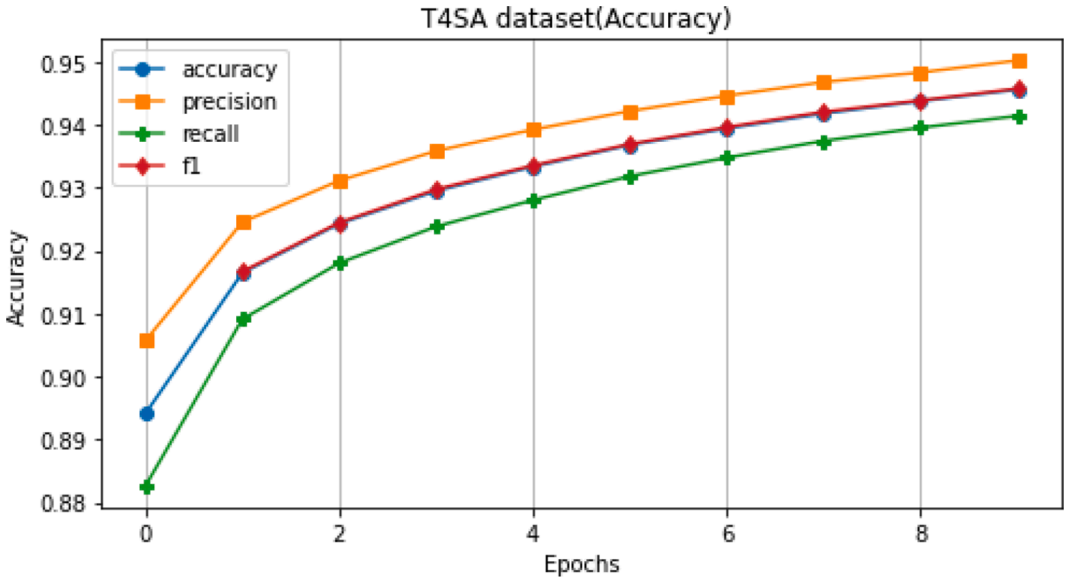

Fig. 13. Accuracy, precision, recall, and F1 on the T4SA dataset (Text).

Table 10
Comparing the results of different models on the MVSA dataset.

F1-score	Accuracy	Approach	Method
0.6419	0.6630	Multi-CNN	(Cai & Xia, 2015)
0.6811	0.6886	MultiSentiNet	(Xu & Multisentinet, 2017)
0.7789	0.7526	DMLANet	(Yadav & Vishwakarma, 2020)
0.7298	0.7298	MVAN-M	(Yang, Feng, Wang, & Zhang, 2021)
0.9366	0.9348	DSET -VGG16- WCNNE and sentence embedding with BERT	The proposed method

Table 11
Comparing the results of different models on the T4SA dataset.

F1-score	Accuracy	Approach	Method
-	0.506	VGG-T4SA FT-F	(Vadicamo et al., 2017)
-	0.513	VGG-T4SA FT-A	(Vadicamo et al., 2017)
-	0.783	CNN	(You et al., 2015)
-	0.773	PCNN	(You et al., 2015)
-	0.5068	Pre-trained CNN (MobileNet v1)	(Ragusa et al., 2020)
-	0.6042	ResNet18, ResNet50, and ResNet152	(Gaspar & Alexandre, 2019)
0.9680	0.9687	Attention-based Bidirectional CNN-RNN	(Basiri et al., 2021)
0.9647	0.9689	DSET -VGG16- WCNNE and sentence embedding with BERT	The proposed method

used as test data.

Tables 10 and 11 show the comparisons of the results of the proposed hybrid model of this paper with the other models on the MVSA and T4SA datasets. In the proposed model of (Cai & Xia, 2015), the textual and visual features of the tweets are learned by two separate CNNs, combined as inputs of another CNN to determine the internal relationship between the texts and the images. In the proposed model of (Xu & Multisentinet, 2017), objects and scenes are extracted from the image and then an LSTM attention model is created to integrate text words with these visual semantic features. The MSA approach of (Yang, Feng, Wang, & Zhang, 2021) is based on the multi-view attentional network (MVAN), in which, a memory network is used, continually updated to take the deep semantic features of image-text. The model includes 3 steps, namely feature mapping, interactive learning, and feature fusion. Self-attention is used for classification in (Yadav & Vishwakarma, 2020), the model of which applies semantic attention to extract the text features related to the bi-attentive image features. The three-step MSA method of (Gaspar & Alexandre, 2019) was evaluated on T4SA. The sentiment of each modal (image and text) is separately analyzed, and the outputs are fused using a weighting method. A hybrid CNN-RNN model was proposed in (Basiri et al., 2021) using an attention mechanism called ABCDM. The polarity findings of the document-level sentiment analysis was focused in this method. The GloVe model was used for word embedding, and bidirectional GRU, bidirectional LSTM, attention mechanism, and CNN were utilized for extracting local features. A hardware-friendly model for image sentiment analysis was proposed in (Ragusa, Gianoglio, Zunino, & Gastaldo, 2020). This model was supported by a CNN using

Table 12
F1-Score of the proposed method for classifying text and image sentiments on the MVSA dataset.

Approach	F1-score (Negative)	F1-score (Neutral)	F1-score (Positive)
Fine-tuning VGG16	0.9091	0.7405	0.5924
WCNNE and Sentence embedding with BERT	0.8935	0.8954	0.8950
Extended Dempster Shafer (Yager)	0.9687	0.9231	0.9105

Table 13
F1-Score of the proposed method for classifying text and image sentiments on the T4SA dataset.

Approach	F1-score (Negative)	F1-score (Neutral)	F1-score (Positive)
Fine-tuning VGG16	0.9317	0.8686	0.8484
WCNNE and Sentence embedding with BERT	0.8862	0.9026	0.9060
Extended Dempster Shafer (Yager)	0.9778	0.9645	0.9627

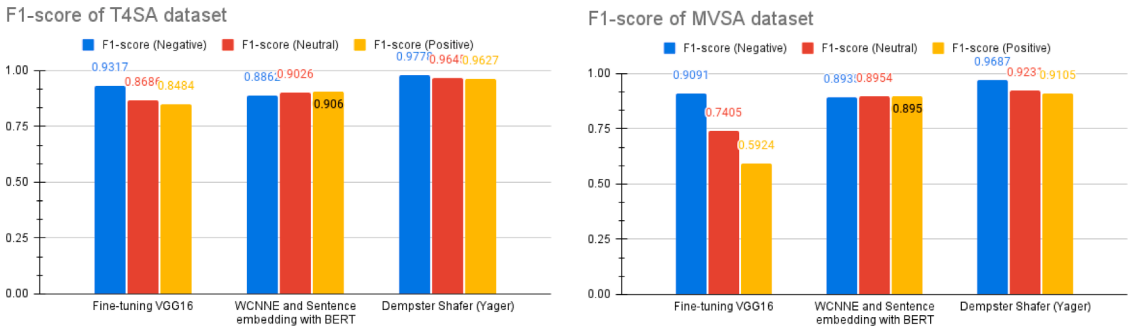

Fig. 14. Comparison of the F1- Score obtained for each polarity in the classifiers and extended Depmster-Shafer (Yager)

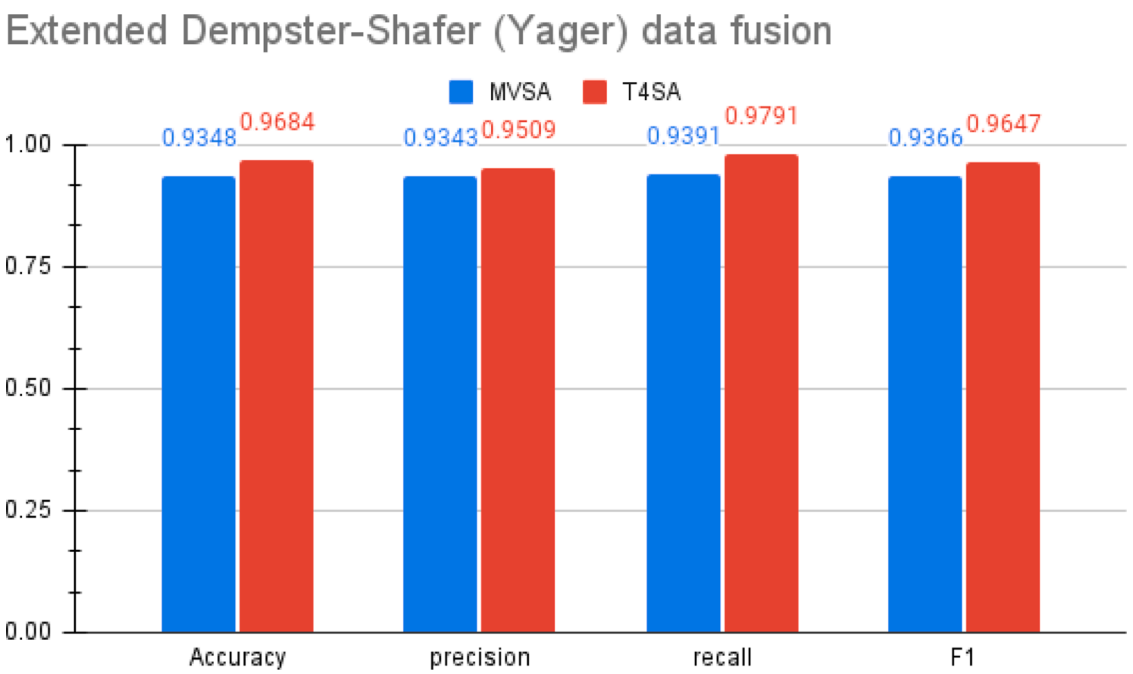

Fig. 15. Comparison of the evaluation results obtained on MVSA and T4SA datasets.

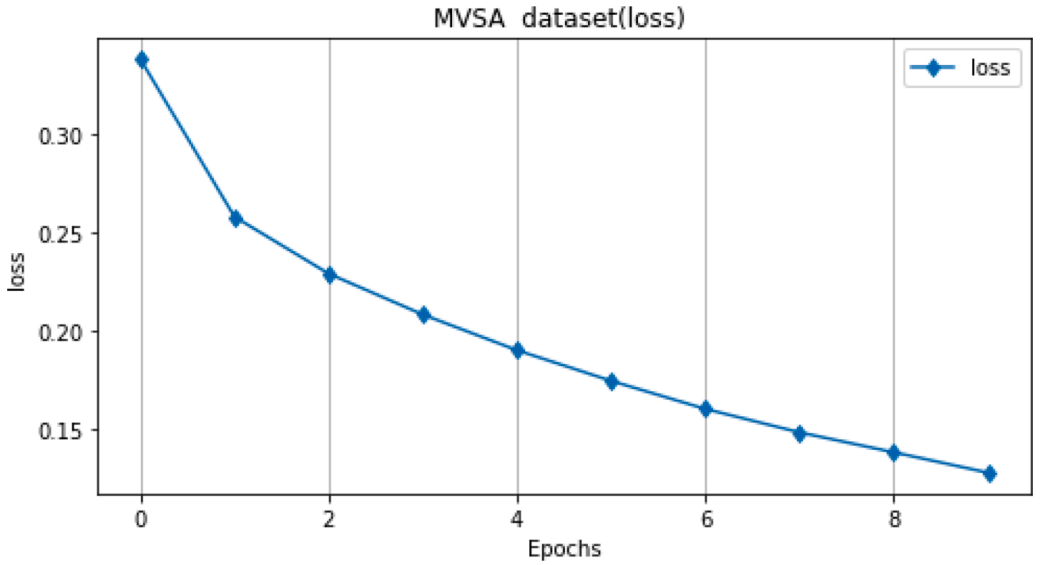

Fig. 16. Loss on the MVSA dataset (Text)

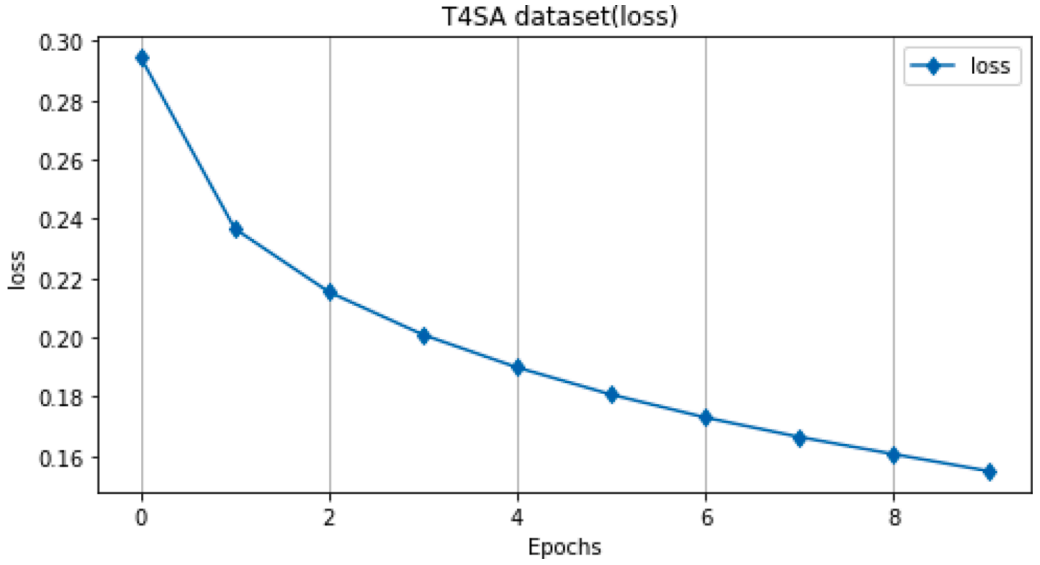

Fig. 17. Loss on the T4SA dataset (Text)

depth-wise separable convolution (DSC) and weight cutting. As shown in Tables 10 and 11, the proposed method of this paper has proper accuracy and F1 score compared to the previous methods Tables 12. and 13 shows F1-Score for every polarity on the classifiers and fusion final decision in MVSA and T4SA datasets Fig. 14. shows the F1-Score of the polarities Fig. 15. shows a comparison between the results obtained for MVSA and T4SA datasets.

5. Error analysis

In this study, the advantages of ensemble learning and transfer learning are utilized to determine the final label of polarity by integrating the outputs at the decision level. To analyze the error of the techniques used based on the experiences gained in the implementation, paying attention to the following points can prevent the failure of the algorithm in AdaBoost:

1 Weak learners should be linear and non-linear basic models should not be used to reduce the probability of overfitting.
2 AdaBoost technique is sensitive to noisy data and cannot lead to appropriate results for noisy data.
3 Weak learning models should not be too weak or too strong.

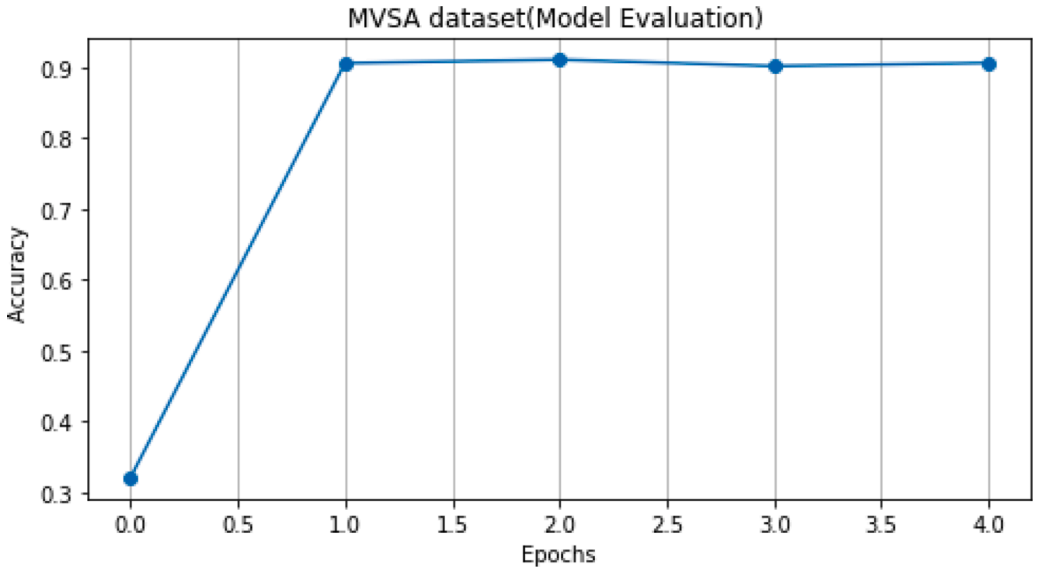

Fig. 18. Evaluation of the proposed model on MVSA dataset (Text)

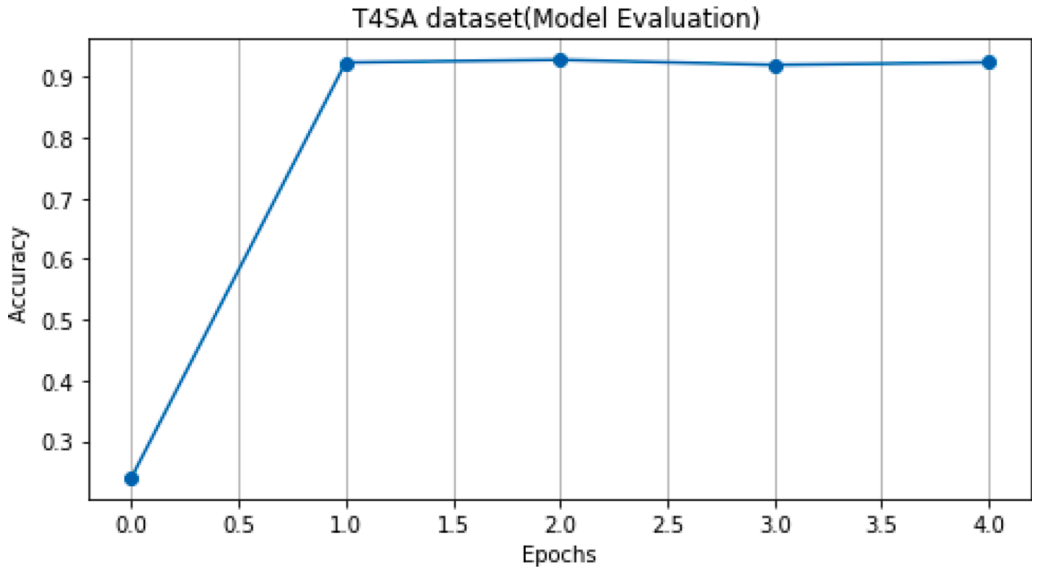

Fig. 19. Evaluation of the proposed model on T4SA dataset (Text)

The VGG16 model, as one of the best models in the classification, has a classification error of 7.32% Fig. 10. shows the amount of the loss in classifying images on the B-T4SA dataset. The classification is performed with the hyper parameter epochs = 10 to classify images, in which, the loss rate decreases and the accuracy rate increases in each epoch. In the first epoch, the loss rate is 0.5489 and the accuracy rate is 0.6340. In the last epoch, the loss rate is 0.2285 and the accuracy rate was 0.85. In the MVSA dataset, the images are classified with different configurations. As shown in Fig. 11, the accuracy rate is 0.84 and the loss rate is 0.2932in epoch 10 Figs. 16 and 17. show the amount of loss in the text section of MVSA and T4SA datasets, and Figs. 18 and 19 show the evaluation of the proposed models on the text section of the datasets.

The confusion matrix can be used to understand the performance of machine learning classification. The negative, neutral and positive polarities are represented in the confusion matrix with 0, 1 and 2, respectively. The true- normalized and pred-normalized confusion matrices for text and image on the datasets MVSA and T4SA are separately shown in Figs. 20–23. For the text section of the MVSA dataset, as shown in Fig. 20, the proposed model performed better in detecting neutral polarity. It has also obtained proper accuracy in detecting positive and negative polarities. For the text section of theT4SA dataset, as shown in Fig. 21, the proposed model performed better in detecting neutral polarity. As shown in Fig. 21, the model can correctly predict 96% and 90% of neutral and positive data on T4SA text dataset, respectively Figs. 22. and 23 show the confusion matrices of the image sections of the datasets. As

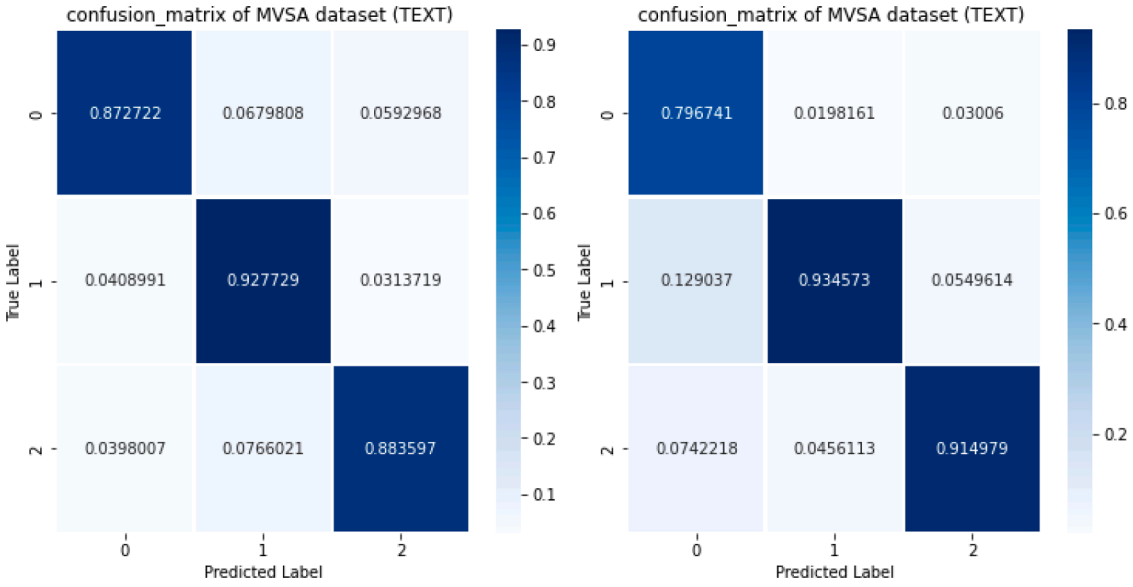

Fig. 20. Confusion matrix of the text section of the MVSA (left: true-normalized, right: pred-normalized)

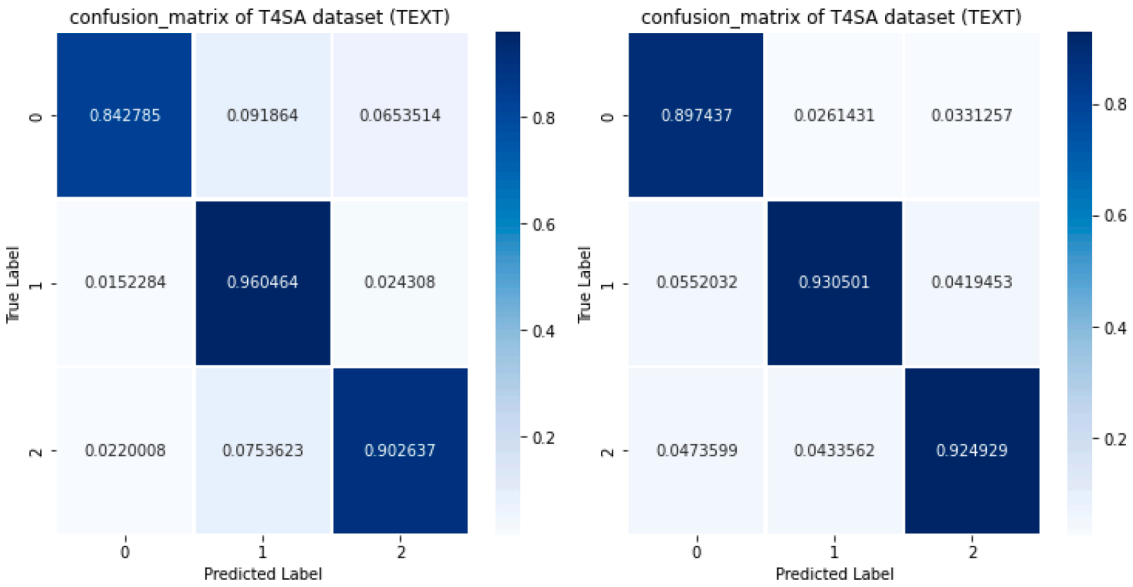

Fig. 21. Confusion matrix of the text section of the T4SA (left: true-normalized, right: pred-normalized)

shown in Fig. 22, the proposed model performed better in detecting negative polarities in classifying the images of the B-T4SA dataset. Finally, Fig. 23 shows the better performance of the proposed model in detecting negative polarity in the true-normalized mode and positive polarity in the pred-normalized mode

6. Conclusions

In this study, a hybrid approach was proposed based on transfer ensemble learning using weighted convolutional neural networks to classify multimodal sentiments in social networks. In the proposed model, the images contained in the message were classified using transfer learning and fine-tuned the pre-trained VGG16 model. The Mask-RCNN model was also utilized to detect objects in images and convert them to text to improve polarity detection using extracted emotions from images. The pre-trained BERT model was used to extract the features and words embedding of the texts extracted from the images and the image captions. The outputs of the BERT model were then used as the inputs of weighted CNNs to extract higher-level features and classify texts. In the proposed AdaBoostCNN, the CNN networks were used as primary estimator with the hyper-parameters. The transfer learning technique used in the proposed

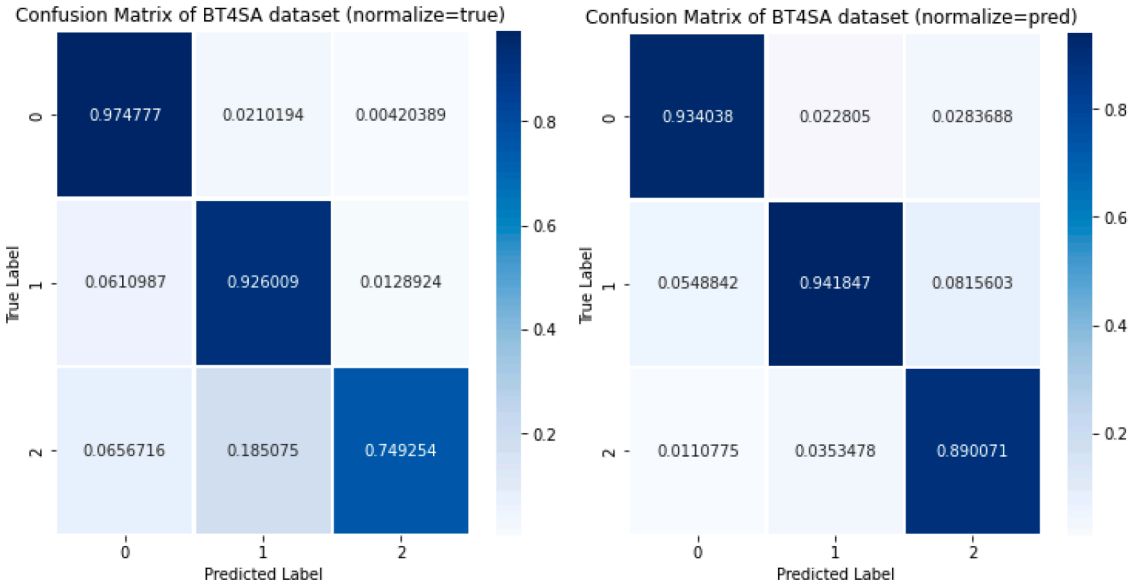

Fig. 22. Confusion matrix of the image subset B-T4SA (left: true-normalized, right: pred-normalized)

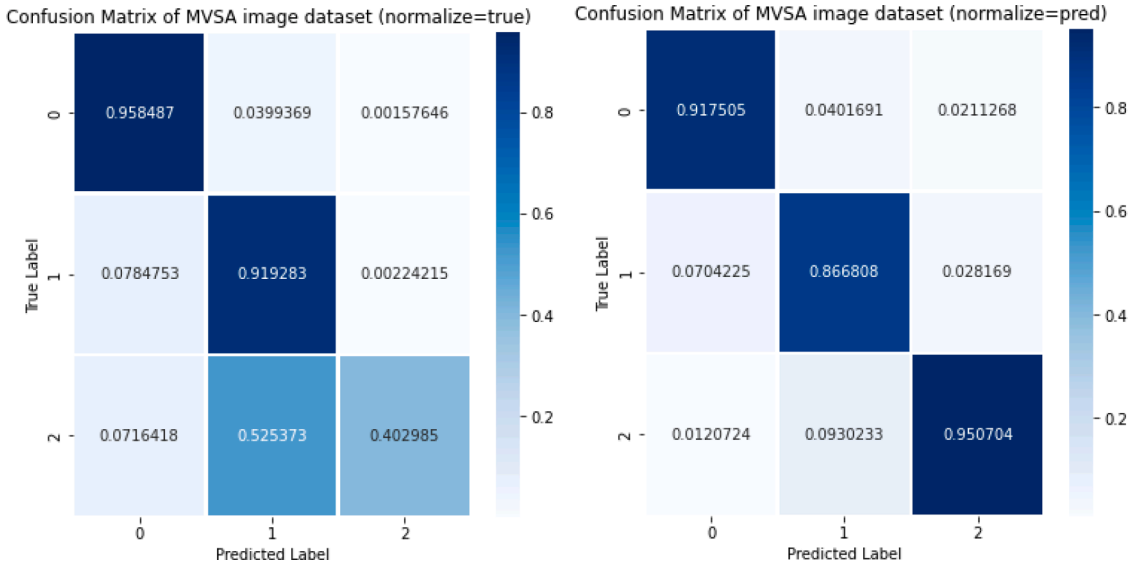

Fig. 23. Confusion matrix of the image section of the MVSA (left: true-normalized, right: pred-normalized)

model to save resources, to reduce training time, and to improve accuracy. The improved Dempster–Shafer (Yager) theory was finally used for fusion at the decision level. Two decision-makers were used for text and image, and the decisions of the classifiers were fused at the decision level. Two multimodal datasets, namely MVSA and T4SA were used for empirical evaluations. Experimental results showed the superiority of the proposed method compared to the other methods in terms of accuracy, recall, and precision. The CNN estimators that make better predictions can be scored and ranked in future works. Using other boosting and fusion approaches can be also considered as other suggestions for future studies.

We propose a new hybrid transfer ensemble learning method based on weighted convolutional neural networks for multimodal sentiment analysis using extended Dempster–Shafer theory.

We used the pre-trained VGG16 network to extract visual features and fine-tune them to our CNN layers for image classification.

To determine the objects in the images and convert them to text, we use the Mask-RCNN model to improve the results of emotion analysis.

CRediT authorship contribution statement

Alireza Ghorbanali: Software, Writing – original draft, Methodology, Validation, Investigation. **Mohammad Karim Sohrabi:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Farzin Yaghmaee:** Conceptualization, Validation, Investigation.

References

Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Systems with Applications, 167*, Article 114324.

Agarwal, N., Sondhi, A., Chopra, K., & Singh, G. (2021). Transfer learning: Survey and classification. *Smart Innovations in Communication and Computational Sciences* (pp. 145–155). Springer.

Ahmed, B., Gulliver, T. A., & alZahir, S. (2020). Image splicing detection using mask-RCNN. *Signal, Image and Video Processing, 14*(7), 1035–1042.

Alayba, A. M., Palade, V., & England, M. (2018). Iqbal R A combined CNN and LSTM model for arabic sentiment analysis. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 179–191). Springer.

Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020). Ensemble neural network approach detecting pain intensity from facial expressions. *Artificial Intelligence in Medicine, 109*, Article 101954.

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems, 115*, 279–294.

Bawa, V. S., & Kumar, V. (2019). Emotional sentiment analysis for a group of people based on transfer learning with a multi-modal system. *Neural Computing and Applications, 31*(12), 9061–9072.

Behera, R. K., Jena, M., Rath, S. K., & Misra, S. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management, 58*(1), Article 102435.

Belhaj, A. F., Elraies, K. A., Alnarabiji, M. S., Kareem, F. A. A., Shuhli, J. A., Mahmood, S. M., & Belhaj, H. (2021). Experimental investigation, binary modelling and artificial neural network prediction of surfactant adsorption for enhanced oil recovery application. *Chemical Engineering Journal, 406*, Article 127081.

Briskilal, J., & Subalalitha, C. (2022). An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. *Information Processing & Management, 59*(1), Article 102756.

Cai, G., & Xia, B. (2015). Convolutional neural networks for multimedia sentiment analysis. *Natural language processing and Chinese computing* (pp. 159–167). Springer.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing, 97*, Article 106754.

Chatterjee, D. P., Mukhopadhyay, S., Goswami, S., & Panigrahi, P. K. (2021). Efficacy of oversampling over machine learning algorithms in case of sentiment analysis. *Data Management, Analytics and Innovation* (pp. 247–260). Springer.

Chen, L.-C., Lopes, R. G., Cheng, B., Collins, M. D., Cubuk, E. D., Zoph, B., Adam, H., & Shlens, J. (2020). Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision* (pp. 695–714). Springer.

Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., & Zadeh, A. (2017). Morency L-P Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 163–171).

Chen, Y., Wang, Y., Gu, Y., He, X., Ghamisi, P., & Jia, X. (2019). Deep learning ensemble for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(6), 1882–1897.

Chen, Y.-F., & Huang, S.-H. (2021). Sentiment-influenced trading system based on multimodal deep reinforcement learning. *Applied Soft Computing, 112*, Article 107788.

Chen, Y.-L., Yeh, Y.-H., & Ma, M.-R. (2021). A movie recommendation method based on users' positive and negative profiles. *Information Processing & Management, 58*(3), Article 102531.

Cheng, Y., Yao, L., Xiang, G., Zhang, G., Tang, T., & Zhong, L. (2020). Text sentiment orientation analysis based on multi-channel CNN and bidirectional GRU with attention mechanism. *IEEE Access, 8*, 134964–134975.

Dashtipour, K., Ieracitano, C., Morabito, F. C., Raza, A., & Hussain, A. (2021). An ensemble based classification approach for persian sentiment analysis. *Progresses in Artificial Intelligence and Neural Systems* (pp. 207–215). Springer.

Dempster, A. P. (1968). Upper and lower probabilities generated by a random closed interval. *The Annals of Mathematical Statistics, 39*(3), 957–966.

Devulapalli, S., Potti, A., Krishnan, R., & Khan, M. S. (2021). Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques. In *Materials Today: Proceedings*.

Dhamayanthi, N., & Lavanya, B. (2021). Sentiment Analysis Framework for E-Commerce Reviews Using Ensemble Machine Learning Algorithms. *Data Engineering and Intelligent Computing* (pp. 359–367). Springer.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science, 14*(2), 241–258.

Duan, X., Ying, S., Yuan, W., Cheng, H., & Yin, X. (2021). QLLog: A log anomaly detection method based on Q-learning algorithm. *Information Processing & Management, 58*(3), Article 102540.

El-Sappagh, S., Saleh, H., Sahal, R., Abuhmed, T., Islam, S. R., Ali, F., & Amer, E. (2021). Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data. *Future Generation Computer Systems, 115*, 680–699.

Frazao, X., & Alexandre, L. A. (2014). Weighted convolutional neural network ensemble. In *Iberoamerican Congress on Pattern Recognition* (pp. 674–681). Springer.

Gaspar, A., & Alexandre, L. A. (2019). A multimodal approach to image sentiment analysis. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 302–309). Springer.

Gepperth, A., & Karaoguz, C. (2016). A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation, 8*(5), 924–934.

Gkoumas, D., Li, Q., Lioma, C., Yu, Y., & Song, D. (2021). What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion, 66*, 184–197.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity, 51*(6), 2623–2646.

Hazarika, D., Poria, S., Zimmermann, R., & Mihalcea, R. (2021). Conversational transfer learning for emotion recognition. *Information Fusion, 65*, 1–12.

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems, 167*, 26–37.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154.

Huddar, M. G., & Sannakki, S. S. (2018). Rajpurohit VS An ensemble approach to utterance level multimodal sentiment analysis. In:. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 145–150) IEEE.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification. *International Journal of Multimedia Information Retrieval, 9*(2), 103–112.

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2021). Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimedia Tools and Applications, 80*(9), 13059–13076.

Ishaq, A., Asghar, S., & Gillani, S. A. (2020). Aspect-based sentiment analysis using a hybridized approach based on CNN and GA. *IEEE Access, 8*, 135499–135512.

Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications, 178*, Article 115019.

Kawana, Y., Ukita, N., Huang, J.-B., & Yang, M.-H. (2018). Ensemble convolutional neural networks for pose estimation. *Computer Vision and Image Understanding, 169*, 62–74.
Kazmaier, J., & van Vuuren, J. H. (2022). The power of ensemble learning in sentiment analysis. *Expert Systems with Applications, 187*, Article 115819.
Kim, Y. (2014). Convolutional neural networks for sentence classification. *CoRR abs/1408.5882 (2014)*. doi: https://doi.org/10.48550/arXiv.1408.5882. arXiv preprint arXiv:14085882.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.
Kumar, A., Srinivasan, K., Cheng, W.-H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management, 57*(1), Article 102141.
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
Li, H., Chen, Q., Zhong, Z., Gong, R., & Han, G. (2022). E-word of mouth sentiment analysis for user behavior studies. *Information Processing & Management, 59*(1), Article 102784.
Li, S., Shi, W., Wang, J., & Zhou, H. (2021). A deep learning-based approach to constructing a domain sentiment lexicon: A case study in financial distress prediction. *Information Processing & Management, 58*(5), Article 102673.
Li, Y., Wei, B., Liu, Y., Yao, L., Chen, H., Yu, J., & Zhu, W. (2018). Incorporating knowledge into neural network for text representation. *Expert Systems with Applications, 96*, 103–114.
Li, Y., Zhang, K., Wang, J., & Gao, X. (2021). A cognitive brain model for multimodal sentiment analysis based on attention neural networks. *Neurocomputing, 430*, 159–173.
Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11–26.
Liu, Y., Li, F., & Ji, D. (2021). Aspect-Based Pair-Wise Opinion Generation in Chinese automotive reviews: Design of the task, dataset and model. *Information Processing & Management, 58*(6), Article 102729.
Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. (2018). Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems, 161*, 124–133.
Marrese-Taylor, E., Balazs, J. A., & Matsuo, Y. (2017). Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. *arXiv preprint*. arXiv: 170802420.
Meškelė, D., & Frasincar, F. (2020). ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management, 57*(3), Article 102211.
Mliki, H., Bouhlel, F., & Hammami, M. (2020). Human activity recognition from UAV-captured video sequences. *Pattern Recognition, 100*, Article 107140.
Morency, L.-. P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces* (pp. 169–176).
Niu, T., Zhu, S., Pang, L., & El Saddik. (2016). A Sentiment analysis on multi-view social data. In:. In *International Conference on Multimedia Modeling* (pp. 15–27) Springer.
Ombabi, A. H., Ouarda, W., & Alimi, A. M. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining, 10*(1), 1–13.
Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications, 57*, 232–247.
Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management, 53*(4), 814–833.
Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management, 53*(4), 764–779.
Pavel, M. I., Razzak, R., Sengupta, K., Niloy, M. D. K., Muqith, M. B., & Tan, S. Y. (2021). Toxic comment classification implementing CNN combining word embedding technique. *Inventive Computation and Information Technologies* (pp. 897–909). Springer.
Poria, S., Cambria, E., & Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods In Natural Language Processing* (pp. 2539–2544).
Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing, 261*, 217–230.
Qu, Z., Mei, J., Liu, L., & Zhou, D.-Y. (2020). Crack detection of concrete pavement with cross-entropy loss function and improved VGG16 network model. *IEEE Access, 8*, 54564–54573.
Qureshi, A. H., Nakamura, Y., Yoshikawa, Y., & Ishiguro, H. (2016). Robot gains social intelligence through multimodal deep reinforcement learning. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)* (pp. 745–751). IEEE.
Ragusa, E., Gianoglio, C., Zunino, R., & Gastaldo, P. (2020). Image polarity detection on resource-constrained devices. *IEEE Intelligent Systems, 35*(6), 50–57.
Rao, G., Huang, W., Feng, Z., & Cong, Q. (2018). LSTM with sentence representations for document-level sentiment classification. *Neurocomputing, 308*, 49–57.
Rezende, E., Ruppert, G., Carvalho, T., Theophilo, A., Ramos, F., & de Geus, P. (2018). Malicious software classification using VGG16 deep neural network's bottleneck features. *Information Technology-New Generations* (pp. 51–59). Springer.
Salunke, V., & Panicker, S. S. (2021). Image sentiment analysis using deep learning. *Inventive Communication and Computational Technologies* (pp. 143–153). Springer.
Sangeetha, K., & Prabha, D. (2021). Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for LSTM. *Journal of Ambient Intelligence and Humanized Computing, 12*(3), 4117–4126.
Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.
Sharma, A. K., Chaurasia, S., & Srivastava, D. K. (2020). Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2Vec. *Procedia Computer Science, 167*, 1139–1147.
Sharmin, S., & Chakma, D. (2021). Attention-based convolutional neural network for Bangla sentiment analysis. *Ai & Society, 36*(1), 381–396.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*. arXiv:14091556.
Smetanin, S., & Komarov, M. (2021). Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management, 58*(3), Article 102484.
Sridharan, K., & Komarasamy, G. (2020). Sentiment classification using harmony random forest and harmony gradient boosting machine. *Soft Computing, 24*(10), 7451–7458.
Stappen, L., Baird, A., Cambria, E., & Schuller, B. W. (2021). Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems, 36*(2), 88–95.
Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., & Falchi, F. (2017). Tesconi M Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 308–317).
Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. *IEEE Access, 8*, 138162–138169.
Xi, D., Xu, W., Chen, R., Zhou, Y., & Yang, Z. (2021). Sending or not? A multimodal framework for Danmaku comment prediction. *Information Processing & Management, 58*(6), Article 102687.
Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., & López, A. M. (2022). Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems, 23*(1), 537–547.
Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: A survey. *Artificial Intelligence Review, 50*(1), 49–73.
Xu, N., & Mao, W. (2017). Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2399–2402).
Yadav, A., & Vishwakarma, D. K. (2020). A Deep Multi-Level Attentive network for Multimodal Sentiment Analysis. *arXiv preprint arXiv:201208256*.
Yadav, A., & Vishwakarma, D. K. (2020). A unified framework of deep networks for genre classification using movie trailer. *Applied Soft Computing, 96*, Article 106624.

Yager, R. R. (1986). Arithmetic and other operations on Dempster-Shafer structures. *International Journal of Man-Machine Studies, 25*(4), 357–366.
Yang, X., Feng, S., Wang, D., & Zhang, Y. (2021). Image-text multimodal emotion classification via Multi-view attentional network. *IEEE Transactions on Multimedia, 23*, 4014–4026.
You, Q., Luo, J., Jin, H., & Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*.
Zhang, D., Li, S., Zhu, Q., & Zhou, G. (2019). Modeling the clause-level structure to multimodal sentiment analysis via reinforcement learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 730–735). IEEE.
Zhang, T., Huang, M., & Zhao, L. (2018). Learning structured representation for text classification via reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
Zhang, Y., Song, D., Zhang, P., Wang, P., Li, J., Li, X., & Wang, B. (2018). A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science, 752*, 21–40.
Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., & Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management, 56*(6), Article 102097.
Zhou, Z.-H. (2021). Ensemble learning. *Machine Learning* (pp. 181–210). Springer.
Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76.