Multimedia Tools and Applications (2021) 80:6871–6910
https://doi.org/10.1007/s11042-020-10037-x

Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks

Jitendra V. Tembhurne[1] · Tausif Diwan[1]

Received: 9 June 2020 / Revised: 31 August 2020 / Accepted: 6 October 2020 /
Published online: 22 October 2020

Abstract

Social networking platforms have witnessed tremendous growth of textual, visual, audio, and mix-mode contents for expressing the views or opinions. Henceforth, Sentiment Analysis (SA) and Emotion Detection (ED) of various social networking posts, blogs, and conversation are very useful and informative for mining the right opinions on different issues, entities, or aspects. The various statistical and probabilistic models based on lexical and machine learning approaches have been employed for these tasks. The emphasis was given to the improvement in the contemporary tools, techniques, models, and approaches, are reflected in majority of the literature. With the recent developments in deep neural networks, various deep learning models are being heavily experimented for the accuracy enhancement in the aforementioned tasks. Recurrent Neural Network (RNN) and its architectural variants such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) comprise an important category of deep neural networks, basically adapted for features extraction in the temporal and sequential inputs. Input to SA and related tasks may be visual, textual, audio, or any combination of these, consisting of an inherent sequentially, we critically investigate the role of sequential deep neural networks in sentiment analysis of multimodal data. Specifically, we present an extensive review over the applicability, challenges, issues, and approaches for textual, visual, and multimodal SA using RNN and its architectural variants.

Keywords Sentiment analysis · Emotion detection · Deep learning · Recurrent neural network · Long short term memory · Gated recurrent unit

✉ Jitendra V. Tembhurne
jitendra.tembhurne@cse.iiitn.ac.in

Tausif Diwan
tausif.diwan@cse.iiitn.ac.in

[1] Department of Computer Science & Engineering, Indian Institute of Information Technology, Nagpur, India

1 Introduction

Now a day, the internet is the integral part of society includes people, organizations, businesses, industries, etc. This is possible by the tremendous growth in communication media and underlying technology such as 4G and 5G. This leverages the availability of communication medium (Internet) to the wide spectrum of applications such as e-commerce, online banking, stock market, social media, etc. Hence, it is observed that the involvement of people increased on internet for various activities specifically online shopping, social networking, and blogs posting, etc. wherein people are engaged in expressing their views and opinions on certain entities or issues. This leads to the development and implementation of automatic recommendation system where users play a vital role by giving their feedback or opinion. To capture the feedback, views, or opinions from the people; various online forums, social networks, and blogs are offered to conduct the discussion or survey on the topic of interest. In these events, feelings, attitudes, views, and opinions are extracted to analyze the conduct of people, expressed as a *sentiment*. The analysis of sentiment is the important, very well researched, and a challenging task in the field of natural language processing (NLP).

Sentiment Analysis can be defined as a broad range covering various subtasks and further categorization under SA. Specifically, it can be defined as a collective process of identifying the sentiment, its granularity i.e. coarse-grained or fine-grained, and analysis of its pros/cons on various targeted entities such as product, movie, sports, politics, etc. The same has been presented in the Fig. 1 to visualize a sub-categorization of SA. However, Emotion detection, closely related to SA, extracts the inherent emotions such as *joy, sadness, anger, fear, trust, disgust, surprise,* and *anticipation* associated with the available data. Generally, an entity or object may pursue several aspects or attributes and different sentiments may be associated with each of these aspects. For example, "*the canvas of shoes is damaged but the sole quality is awesome*", here shoes is identified as *object* with aspect terms *canvas* and *sole*. This problem is known as aspect level sentiment analysis consisting of three major steps; extraction of aspect terms, finding the associated sentiment for each of the aspect terms and lastly to draw the overall sentiment for the object in question. Further, Sarcasm Detection (SD) is the extraction and analysis of inherent negativity with the associated data. Table 1 summarizes SA and related tasks along with its important characteristics.

Recently, people's active involvement on Twitter, Facebook, and Instagram has increased [15, 103]. The opinions related to various topics are freely expressed by the people in the form

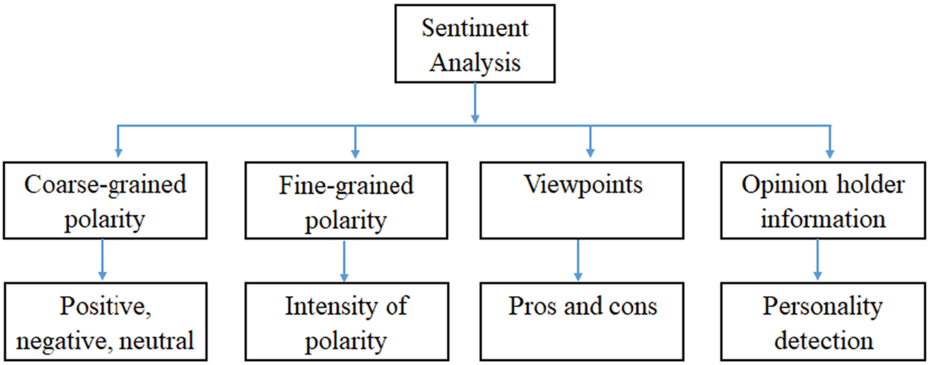

Fig. 1 Sub-categorization of SA

Table 1 SA and Related Tasks

Sr. No.	Sentiment Analysis or Related Tasks	Characteristics
1	Sentiment Analysis	Extraction and Analysis of Positivity, Neutrality, or Negativity about any *aspect*, *entity*, or *issue*.
2	Emotion Detection	Extraction and Analysis of inherent emotion such as *joy, sadness, anger, fear, trust, disgust, surprise, anticipation,* or *pride*.
3	Sarcasm Detection	Identification of extremely hidden negativity or positivity such as *remark*, *reaction*, *taunt*, *cutting expression*, and *mostly hidden meanings*

of text, images, videos, etc. Hence, data is generated in large capacity on Twitter, Facebook, Instagram, and on other social networking sites. So, various researchers are attracted to analyze this data to investigate the SA and related tasks. Due to this, SA is widely researched and applied on a wide range of domains to compute accurate *sentiments* and corresponding *emotions*.

1.1 Applications of SA

The SA is applied in **products reviews** [40], movie review [93], stock market prediction [152], and opinion polls [14], etc. The product review is very helpful in understanding the features and shortcomings in the product purchased by the customers. So, reviewing various posts related to product facilitated the customer for his/her likeliness to purchase the product and also suggest the company to improve upon the drawbacks in the product. Fang and Zhan [40] investigated SA of product review on *Amazon.com* to address the problem of categorization of sentiment polarity. The sentence- and review-level classification is accomplished by using random forest (RF), Naïve Bayesian (NB), and support vector machine (SVM) [81]. The product reviews mostly influence the credibility of product and decide the future of the products in the market. So, careful analysis of these reviews has become the basic need to achieve potential businesses. Product features and global score identification using data mining on big data text is studied in [74]. The sentiments are analysed for decision making to market the product. In [165], deep learning is adopted for sentiment classification on product review wherein product rating considered as a weak signal. This study helps the customers to get the visibility in terms of buying decision. Recently, combination of deep learning and sentiment lexicon are utilized for SA on product review [148]. Sentiment features are enhanced by sentiment lexicon and leading sentiment features are extracted by convolutional neural network (CNN) along with gated recurrent unit (GRU). Not only product review but credibility of the reviewers also plays a vital role in the product recommendation. In [57], new direction is investigated for the product recommendation to the customer. To do so, the entire profile of reviewers is taken into account to design a model to analyze the sentiments over product review, sentiment confidence, and context; which is utilized to extract the important reviewer features.

Nevertheless, **Stock market** is one of the important financial entities of any nation. The stock market drives the economy of nation wherein people, organization, and government do their investments. Stock market is completely uncertain and needs detailed analysis to make the investment. Subsequently, SA also applied on stock market to predict the sentiments for the investment. A machine learning approach is applied [141] on opinion posts of stock market shared by investors online. To predict the stock sentiment, SA model is employed wherein the feature selection is followed by feature reduction and finally classification of stock opinions is performed using SVM classifier. For the experimentation of stock opinion, Sina Finance platform is utilized. Forecasting stock opinion for investment in a stock using investors' sentiment is explored in [114]. The financial data and news data are pre-processed, features are extracted, and expressed in the form of sentiment index. Eventually, SVM is employed for stock market forecasting of SSE-50 Index and reported high accuracy of 89.93% using sentiment variable. Another machine learning model is developed [76] to predict the trends in the stock market. The effect of political events and people's sentiment are considered to monitor the influence on stock market. For accurate prediction, the situation and sentiment are chosen as the features, experimentation is performed on the data obtained from Yahoo! Finance, Twitter and political events from Wikipedia. Moreover, the behavior of stock is mostly depends on investors' sentiment which has the ability to predict stock returns. The proposal [98] to find the association among stock return and investor sentiment by using SA on social media texts is demonstrated. The effect of investor sentiments has significant influence on unusual returns of stock. Recently, the impact of sentiment and its hidden emotion are used as variables in the prediction of stock movement using NB, SVM and *k*-nearest neighbors (k-NN) [126]. The different variables perform a vital role to predict the stock movement accurately.

1.2 Datasets of SA

Various standard and labeled datasets are found in the literature for the SA determination process. Table 2 highlights various most popular datasets used in SA. The data required in the SA process should be specific, well annotated and labeled, and large in size. The smaller datasets don't perform well in accordance with the deep models. The datasets for SA are characterized by the amount of preprocessing required, format, and number of instances in the datasets (https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research).

Deep Learning is an extension of Machine Learning wherein deep neural networks are employed for feature extraction and analysis from large sized datasets. There are various deep neural network models such as CNNs basically used for automatic feature extraction without any explicit feature engineering, Recurrent Neural networks used for the feature extraction and classification from the temporal or sequential data. Input to the SA and related tasks may be visual, textual, audio, or any combination of these, RNN models seems to be very efficient for solving this set of tasks having an inherent sequential nature of input. In this paper, we present a critical review on SA and related tasks using deep learning with a focus on Recurrent Neural Networks. The followings are the major contributions of this paper:

1 Demonstration of the applicability of SA in various domains.
2 Summarization of state of the arts surveys and reviews on SA using machine learning and deep leasing based approaches.
3 To present and discuss on deep learning and sequence models such as RNN, LSTM, and GRU along with a schematic comparative illustration.
4 Critical and detailed investigation covering the challenges, role, applicability, and approachability of RNNs and its architectural variants in textual, visual, and multimodal SA.

Table 2 Popular Datasets for SA

Category	Datasets	Description
Movie Reviews	IMDB [62]	Stands for Internet Movie Database, largest database consisting movies information and user reviews. It has 83 million registered users as on October 2018 along with their opinions and sentiments for several movies.
	Stanford Sentiment Treebank [139]	Movie reviews in two flavors with 5 classes and 2 classes for fine grained and coarse-grained polarity respectively.
News articles	Thomson Reuters text research collection [88]	News articles about an specific allegations
	NYSK dataset [33]	Large collection of various news stories consisting of 10,421 news articles on sexual assault (XML, text)
	ABC Australia News Corpus	A news corpus of ABC Australia consisting of 1,186,018 news articles from 2013 to 2019 (CSV)
Twitter dataset	Sentiment labeled sentences dataset [79, 104]	This dataset consists of 3000 sentences along with their hand annotated polarity (positive and negative).
	Sentiment 140 [25, 43]	This dataset contains 1,578,627 tweets from 2009 including timestamp, opinion holder information, and its associated polarity (CSV)
	Twitter Dataset for Arabic Sentiment Analysis	2000 Arabic tweets (Text)
Product Review	Amazon Reviews	A huge collection of 82 M reviews approximately on US product review from Amazon.com (Text)

The rest of the paper is organized as follows: Section 2 highlights some of the important surveys and reviews related to SA based on different machine learning and deep learning models. It also presents the advantages and disadvantages of different models for SA. Section 3 presents the detailed discussion of the sequence models viz. simple-RNN, LSTM, and GRU based on their gating mechanism along with a comparative analysis of these models. Section 4 helps in understanding the challenges, issues, and applicability of RNNs in textual SA. The challenges and applications of visual SA using RNNs is illustrated in Section 5. The multimodal SA using RNNs is discussed in Section 6 with focus on challenges and applications of RNNs. Section 7 concludes the paper and provides the future research direction. However, the complete detailed flow of the paper is illustrated using Fig. 2.

2 Previous reviews/surveys on SA

Various survey and reviews based on SA, ED, and SD using machine learning, lexical approaches, and deep learning have been presented by the researchers. In this section, we summarize some of the important review papers that cover all the important aspects for these tasks such as SA challenges, classification of existing SA approaches, deep learning models for SA, classification of deep models, suitability of discriminative models in SA, and handling multimodality in SA.

In [61], **challenges** in the SA and ED are reviewed based on the technical and theoretical aspects. Authors observed and listed various challenges in the identification of sentiments and emotions for different types of inputs such as text, audio, video, and mixed modality. The theoretical challenges includes; negation, domain dependency, bi-polar words, huge lexicon, entity features, NPL overheads, and spam and fake detection. The technical challenges

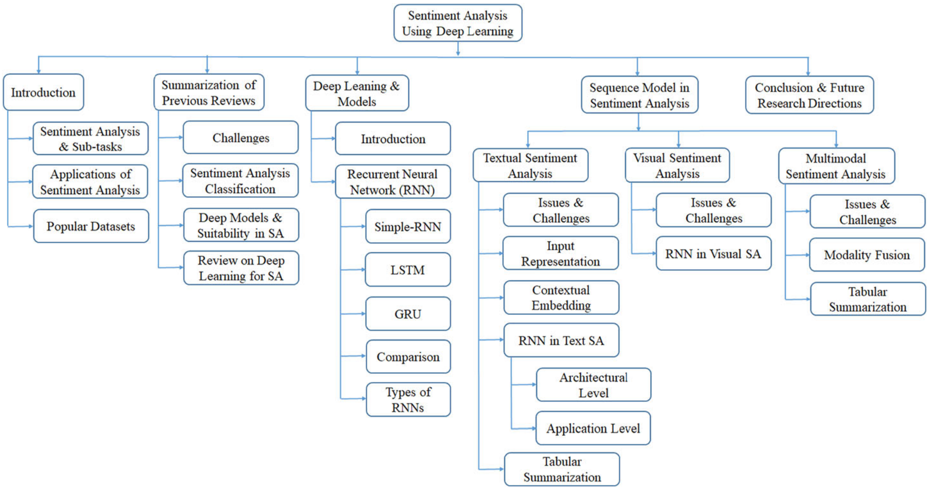
Sentiment Analysis Using Deep Learning
Introduction
Sentiment Analysis & Sub-tasks
Applications of Sentiment Analysis
Popular Datasets
Summarization of Previous Reviews
Challenges
Sentiment Analysis Classification
Deep Models & Suitability in SA
Review on Deep Learning for SA
Deep Leaning & Models
Introduction
Recurrent Neural Network (RNN)
Simple-RNN
LSTM
GRU
Comparison
Types of RNNs
Sequence Model in Sentiment Analysis
Textual Sentiment Analysis
Issues & Challenges
Input Representation
Contextual Embedding
RNN in Text SA
Architectural Level
Application Level
Tabular Summarization
Visual Sentiment Analysis
Issues & Challenges
RNN in Visual SA
Multimodal Sentiment Analysis
Issues & Challenges
Modality Fusion
Tabular Summarization
Conclusion & Future Research Directions

Fig. 2 Organization of the Review

addressed in SA process were extracting features or keywords, word knowledge, bi-polar words, and huge lexicon. These challenges were identified based on the review structure collected for SA/ED i.e. sentiments for structured, semi-structured, and unstructured reviews.

Subsequently, cross domain SA is gaining huge popularity wherein model is trained on the dataset of one domain and performance of the model is evaluated on the dataset of similar or dissimilar domain. For example, a machine learning model is trained for sarcasm detection on movie review tweets and the performance of the same model is evaluated on customer reviews of car buyers. Cross domain SA is basically an indication of good generalization of model across domains. However, this is not the easy task and generally models don’t perform well in cross domain SA. Polarity may be reversed while migrating from one domain to another. The systematic review of various methods, techniques, and approaches adopted in cross domain SA is presented in [3, 72]. The main reason behind cross domain SA is the absence of annotated data in all the domains for SA tasks. Features deviation, polarity reversal, and lexical ambiguity are summarized as the major challenges in the cross domain SA. The difference between cross-domain SA and in-domain SA performance, analysis of role and importance of data representation, and the impact of homogeneity and heterogeneity are the major issues in cross domain SA [166]. The selection and performance of source domain is an extremely important step in cross domain SA. Text similarity features are investigated in cross domain SA for the identification of most suitable source domain while giving the target domain [123]. A precision over 50% is achieved by employing 11 similarity metrics for all the combination of 20 domains in the identification of *k* most suitable source domains.

The **classification** of SA is broadly categorized in lexicon based [34] and machine learning-based approaches [7]. Lexicon based approaches require lexical knowledge i.e. a collection of sentiment words in the respective domain. A sentiment score is associated with each positive and negative word in the sentiment lexicon. Linguistic and domain knowledge become a bottleneck in this approach. Machine learning approaches for SA basically follows; assigning weightages to the extracted features followed by features selection, and then applying an appropriate machine learning model. Generally, machine learning architectures for SA are shallow neural network such as Gaussian mixture model (GMM), hidden Markov model

(HMM), and support vector machine. These models are unable to exploit multiple layers of non-linear features and suffer from high dimensionality and sparsity of the features. Though, these approaches produce comparable results but with the advent of deep learning approaches, considerable better results have been obtained.

The **SA classification** is further fine-tuned and presented in Fig. 3. Though, neural networks are categorized under linear classifier in Fig. 3, the non-linearly is introduced in the neural networks using various non-linear activation functions. For each and every subcategory, algorithmic process for SA and its related tasks are summarized with their pros and cons [99]. Specifically, a summarization of approximately fifty-four articles specifying tasks (SA or ED or transfer learning or building resources or sentiment classification or feature selection), domain orientation, approaches, fine grained or coarse-grained polarity, and datasets is nicely sketched. Similarly, categorization of different methods employed in SA and ED is presented in [35]. The identified methods i.e. machine learning, rule-based, and lexicons based were distinguished with their advantages and disadvantages. Further, comparison of various machine learning methods based on their functionalities and utilities were exemplified. These important machine learning methods are – SVM, N-gram, Naïve Bayes, maximum entropy classifier, k-NN, weighted k-NN, multilingual, and feature driven SA. SVM is one of the important supervised machine learning technique used for classification and regression. It is based on providing a hyperplane that distinguishes the classes from a maximum margin in case of classification problems. N-gram is an important probabilistic machine learning model that predicts a word followed or preceded by a sequence of *N* consecutive words. Eventually, a brief review to classify and compare different methods based on advantages and disadvantages for opinion mining and SA is presented in [52] wherein authors systematically outlined steps and levels of opinion mining. Furthermore, comparison, summarization, and classification of techniques from the literature are discussed based on sentiment classification, aspect extraction, and production and evaluation of summary. For better indulgent, various factors were identified to summarize the different techniques based on their pros and cons. The authors demonstrated that the supervised approaches are best suited for classification and better prediction in SA. Moreover, semi-supervised approaches are also the good candidate for micro-blogs SA.

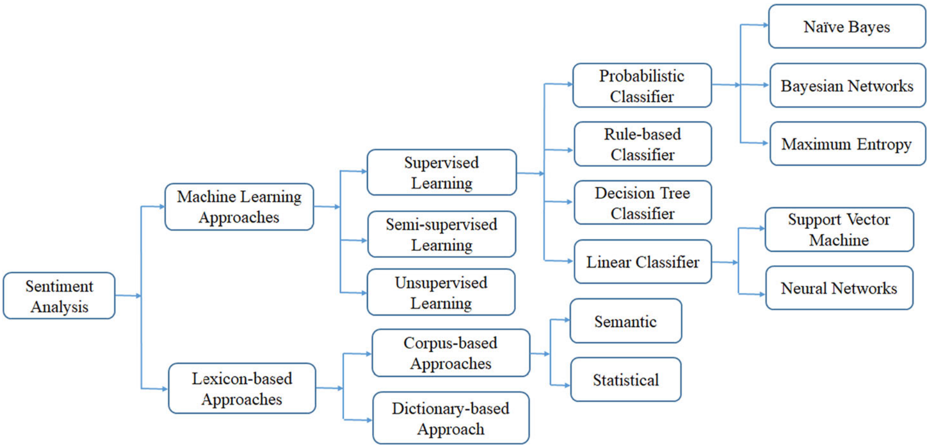

Fig. 3 Classification of SA Techniques

In [31], authors present the architectures, algorithms, and applications to deep learning in the context of NLP. The NLP is a branch of Artificial Intelligence (AI) that describe an ability of a computer program to understand, interpret, and manipulate human languages. Deep neural networks are categorized into three major components viz. *generative*, ***discriminative***, and *hybrid* architectures, and category-wise algorithms and applications were explored. As textual SA is an important application of NLP, understanding the discriminative deep architectures would help us to incorporate them in the betterment of textual SA. Deep generative models are further analyzed to an application side and explored various applications of it in the fields of image processing and information retrieval [144]. In one of these application segments, deep learning models are summarized for three subtask of SA viz. sentiment classification, sentiment extraction, and building sentiment lexicon; majorly focusing on textual input. In the same line, various deep learning models are reviewed for SA and ED observing a better semantic knowledge extraction without explicit feature engineering [128]. It provides an enrich understanding of the different approaches for SA and related processes. These processes are word embedding, sentiment classification, opinion extraction, and lexicon learning. In addition, the implementations of deep neural networks for the beginner are suggested. Moreover, the challenges faced by deep learning methods are highlighted with several fruitful suggestions to overcome the existing challenges.

Subsequently, deep learning based SA review is proposed in [99]. The variety of NLP tasks i.e. sentiment classification, textual analysis, visual analysis, product reviews, and cross-lingual analysis has been discussed. The various Deep learning methods have been reviewed to solve these problems. The targeted deep learning models for this review were Deep Neural Networks (DNN), CNN, Recursive Neural Networks (RecNN), RNN, Deep Belief Networks (DBN), etc. In [162], authors attempted to review the SA by applying various deep learning techniques. Sentence level, document level, aspect level, aspect extraction and categorization, opinion extraction, sentiment composition, opinion holder information extraction, emotion analysis, multimodal SA, and multilingual analysis based research papers have been reviewed on the ground of textual representation and underlying neural network models. Three important datasets viz. *Movie review dataset*, *Sentiment Treebank*, and *Twitter dataset* are tested for SA with deep neural networks and concluded that deep CNN outperforms for binary as well as fine-grained SA [115].

Recently, survey presented by [147] highlighted the popularity of deep learning models and their applicability in SA. The classification of SA and its inferences in deep learning model is conferred. In addition, comparison of different deep learning prototypes based on datasets and its features, prototype applied, and accuracy obtained is analyzed. In [29], comparative study is proposed for SA on the data provided in social networks, exclusively for Facebook or Twitter. The authors reviewed latest articles based on the problem of sentiment polarity in the SA process. The word embedding and term frequency-inverse document frequency (TF-IDF) is applied on DNN, CNN, and RNN for different datasets to visualize the accuracy of deep learning models. Furthermore, a comparative study on sentiment classification is presented by [84]. Three CNNs and five RNNs are experimented to derive important inferences to build suitable model for sentiment classification. Moreover, the character- and word-level input types are considered for the datasets of services and products i.e. total 13 review datasets. The performance of various models is evaluated based on dataset characteristics, input level, and model used. The findings are highlighted as; classification performance is higher for larger datasets, classification of sentiments is better for word-level input as compared to character-level input, and the effects of model complexity based on CNNs and RNNs was observed where RNNs wins the race. Also, the uses of LSTM or GRU increase the performance, and additionally improved performance is achieved by using bidirectional LSTM (Bi-LSTM).

Due to Social media advent, SA is not limited to process texts but sentiments are also derived from the images and videos. People find very easy to express the opinions in the form of images and videos and it is also increasing day-by-day. This corresponds to the involvement of more than one modality in the input data. Subsequently, multimodal SA is explored and researched to determine the opinion or sentiment from mixed modality contents. Involvement of facial and vocal expressions in addition to the textual content is offering a tremendous boost in the performance of SA. In contextual multimodal SA, context plays a vital role in SA determination that may not always be captured through only textual data. In addition, important clues from visual and audio modalities are also utilized. The various difficulties and opportunities for the improvements in the multimodal SA is presented in [75] that also covers the taxonomy wise categorization of the techniques involved in multimodal SA. Authors also demonstrated the categorization of SA process based on emotion and opinion mining. In [41, 125], multimodal SA is reviewed and summarized for vLogs and spoken words, visual-textual inputs, human-machine interaction and human-human interaction. The opportunities and difficulties in multimodal SA are listed by the authors. The majority of reviews are broadly focused on the applications of various deep or machine learning models in SA. However, our review specially focuses on the role, challenges, applicability, and various approaches to tackle these challenges using sequence models such as RNN, LSTM, and GRU in textual, visual and multimodal SA.

The challenge in the multimodal SA is to extract and process individual modalities of the multimodal data. In deep learning, generally CNN and its architectural variants are employed for processing visual features. However, RNN and its architectural successors are being experimented to model sequential data such as videos (frames of images), textual, and audio contents. Feature level fusion and Decision level fusion is also compared extensively for the detection of overall sentiment polarity and extraction of inherent emotion from the multimodal data. Capturing the context from multimodal data and proportion of context dependency on individual modality is another major challenge in multimodal SA because context plays a very important role in the SA determination.

3 Deep learning, models, and performance measures

Deep learning is emerged in 2006 from machine learning in which deep neural networks are architectured for the minimization of the loss or error components. It incorporates representational feature learning and synthesizing features in incremental fashion using multiple layers of neural networks [11, 53, 96]. Convolutional neural network is a deep neural network, employed heavily in computer vision tasks, seems to be a powerful tool for extracting spatial features information from visual inputs. Due to fewer parameters as compared to the standard feedforward fully connected neural networks, CNNs are easier to train and test. It is a powerful computational tool that offers learning in both supervised and unsupervised manner. Recurrent Neural Networks and its architectural successors such as LSTM and GRU are basically employed for the features extraction in the sequential, temporal or time-series data. They play an important role in the contextual features extraction in the case of time-series data, as most of the input streams to the aforementioned NLP tasks are in the form of sequences. In this section, we present and discuss the architectural designs of the different deep learning models such as CNN, RNN, LSTM, and GRU with major emphasis to RNN and its family.

3.1 Convolutional neural networks

Due to scaling inefficacy in fully connected neural networks, CNN is widely adopted to capture the spatial and contextual information with fewer parameters. While dealing with high-dimensional inputs, it is almost impractical to connect a neuron in a given layer to all neurons in the previous layer. Instead, we connect each neuron to only a part of the previous layer. This is the basic philosophy behind the working of CNN model. From the architectural viewpoints such as *convolutional*, *pooling*, and *rectified linear unit* (ReLU) collectively act as a basic transformation unit converting an input volume to an output volume. The spatial extent in the convolution operation is a hyperparameter known as a *receptive-field* or *filter-size*. *Filter-size* that convolves over the input plays an important role in extracting useful features information. Other hyperparameters *depth*, *stride*, and *zero-padding* decide the size of the output volume [10]. Herein, we have not covered CNN in detail as it is beyond the scope of this paper. All the important aspects of the CNN such as architectural details, applications, and recent advances in the CNN are nicely presented in [46].

3.2 Recurrent neural networks

RNN is an important deep neural network designed for feature learning in a sequential, temporal or time series input. RNNs are used to solve many scientific problems with high accuracy and widely used in variety of areas ranging from bioinformatics to stock market prediction. The advancements of RNNs are explored and experimented in due time by various researchers [120]. The applications of RNNs and its architectural variants may be appreciated in various works such as weather forecasting [158], stock market prediction [113], speech recognition [119], object detection [86], character recognition [163], intrusion detection [77], automatic landslide detection [100], time series prediction [19], text classification [80], gene expression [41], micro-blogs [159], biological data handling [87], unstructured text data mining with fault classification [136], video processing such as caption generation [145], and many more.

In sequential or temporal input, data at a time-step has relevance over the data of the preceding time-steps. Prediction at any instant is not only determined by the instantaneous input but depends on past history also. In other words, another dimension i.e. *temporal ordering* is also taken care in all RNN model computation. This philosophy is the backbone of the RNN computation. The architectural design (folded in *left* side and unfolded form in the *right* side) of the simple RNN is illustrated in Fig. 4. Output at each time-step is evaluated and the corresponding hidden state is ingested into the successor as depicted in Fig. 4 wherein x, s, and o represents *input*, *hidden*, and *output nodes* respectively. U, V, and W are the shared weight matrices from *input-to-hidden*, *hidden-to-output*, and between consecutive hidden nodes respectively across all the time-steps. The input at i^{th} time-step is x_i b_s and b_o are the biases for the hidden and output node respectively. The model is generally trained using the backpropagation algorithm, known as backpropagation through time (BPTT) [138] that incorporates the notion of the time/sequence in the underlying gradient descent process.

The gradient is getting diminished for long intervals in the course of back propagation as the smaller derivatives are multiplied using a chain rule, resulting in the negligible weight change for the distant weight matrices. Formally, we are unable to capture the long term dependency across the distant part of the inputs in the plain RNN. This problem is known as vanishing gradient problem and considered as a major bottleneck in the traditional RNN [55].

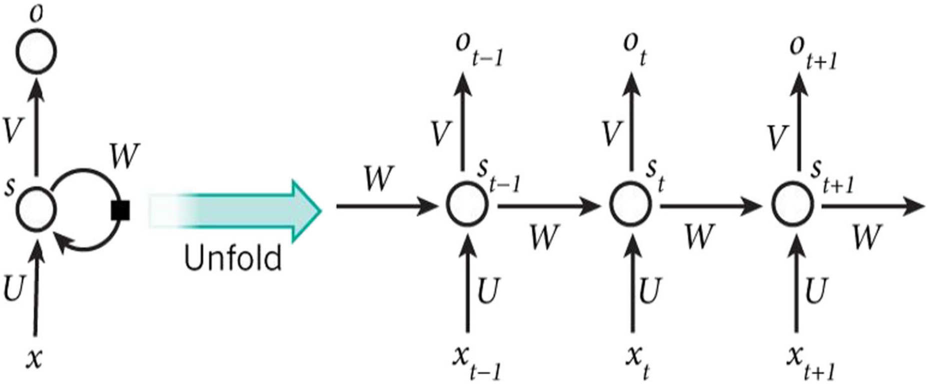

Fig. 4 Unfolding of Recurrent Neural Network Schematic [13]

To better model the long term dependency and to mitigate the effect of vanishing gradient problem, advancements in the architectural designs of RNN are proposed in the form of its architectural variants such as LSTM [106] and GRU [27]. These models leverage the advantages of *gating* mechanism to realize long term dependency very well. The internal gates of LSTM and GRU cells decide the flow of the information in the network. In every cell operation, important information is retained and transmitted further whereas the non-necessary information is blocked. The network learns which information is relevant and should be kept or forgotten during the training phase of the model. This is accomplished by maintaining cell-state information that acts as a conveyer belt, add the important information or remove the non-necessary information as and when needed. Here, *sigmoid* activation plays an important role in distinguishing between the important and useless information as this function squishes between 0 to 1.

Three gates namely *forget*, *input*, and *output gate* are the important pillars of a LSTM cell. The *forget* gate processes the input of current timestep and hidden output from the previous cell. Cell state gets manipulated due to the various gates operations in LSTM cell i.e. information is added, retained, or subsidized. Cell state gets modified by taking into account the *input gate, forget gate* and previous cell state. *Output* gate is responsible for the generation of the hidden state that shall be utilized in the next LSTM cell. Equations corresponding to the LSTM i.e. Eqs. 3–8, subscripts of the weight matrices and biases indicate the *gate's initial*. A lot of architectural improvements are proposed by various researchers depending upon the application requirements and other heuristics. Some of the architectural advancements in the LSTM can be viewed as sentence-state LSTM (S-LSTM) [160], stacked LSTM [70], bidirectional LSTM [45], and multidimensional LSTM [44].

GRU is very much similar model in comparison with the LSTM, adopted as a ramification for the same vanishing gradient problem. This model comes with lesser tensor operations i.e. *reset* gate and *update* gate are employed for the modeling of long-term dependency. *Update* gate is responsible for the collective functioning of the *forget* and *input* gate of a LSTM cell whereas *reset* gate determines the amount of the past information to be kept or forgotten. Comparative analysis of these three RNN models on the ground of cell operations, underlying equations, model complexity, key characteristics, and shortcomings are highlighted in Fig. 5, Eqs. 1–12, and Table 3.

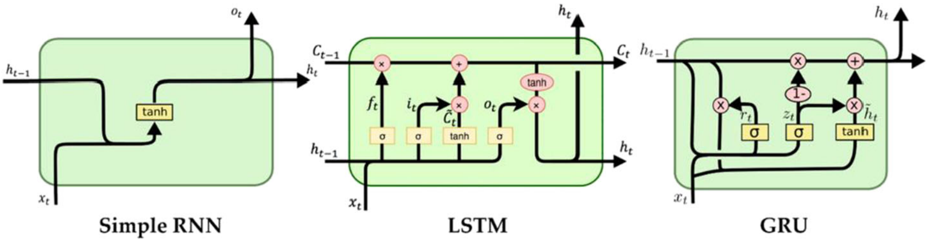

Fig. 5 The indepenent cells of RNN, LSTM, and GRU

The cell equations of simple RNN consist of hidden state (h_t) and output (o_t) are expressed as;

$$h_t = \tanh(W \times h_{t-1} + U \times x_t + b_s) \tag{1}$$

$$o_t = \sigma(V \times h_t + b_o) \tag{2}$$

The LSTM cell equations corresponding to *forget* gate (f_t), *input* gate (i_t), cell state (C_t), *output* gate (o_t), and hidden state (h_t) are represented as follows;

$$f_t = \sigma\left(W_f \times [x_t, h_{t-1}] + b_f\right) \tag{3}$$

$$i_t = \sigma(W_i \times [x_t, h_{t-1}] + b_i) \tag{4}$$

$$\widetilde{C}_t = \tanh(W_c \times [x_t, h_{t-1}] + b_c) \tag{5}$$

$$C_t = C_{t-1} \times f_t + \widetilde{C}_t \times i_t \tag{6}$$

$$o_t = \sigma(W_o \times [x_t, h_{t-1}] + b_o) \tag{7}$$

Table 3 Comparative analysis of RNN and its architectural variants

	Simple RNN	LSTM	GRU
Model Complexity	Low	High	Moderate
Key Characteristics	Easier to train, Less computational Resources	Model long term dependency, Extraction of contextual Information	Model long term dependency, Extraction of contextual Information
Shortcomings	Vanishing gradient problem	High hidden layer complexity	Higher complexity than simple RNN

$$h_t = C_t \times \tanh(o_t) \tag{8}$$

Moreover, the cell equations for GRU consist of *reset* gate (r_t), *update* gate (z_t) and hidden state (h_t) are expressed as;

$$r_t = \sigma(W_r \times [x_t, h_{t-1}] + b_r) \tag{9}$$

$$z_t = \sigma(W_z \times [x_t, h_{t-1}] + b_z) \tag{10}$$

$$\widetilde{h}_t = \tanh(r_t \times [x_t, h_{t-1}] + b_h) \tag{11}$$

$$h_t = z_t \times \widetilde{h}_t + (1 - z_t) \times h_{t-1} \tag{12}$$

Depending on the distribution of input and output across different time-steps, RNNs may be categorized at application level and the same is illustrated in Fig. 6. The figure is self-explanatory and demonstrating all possibilities of input-output distributions in temporal orders along with the examples of each categories such as *textual* and *visual* SA (using videos as inputs i.e. frames of images) belong to the *many-to-one* category as we predict the sentiment only after providing the complete input.

The various **performance measures** are adopted for the evaluation of any machine learning or deep learning models. These performance measures have their specific significance and impact. We will not cover these measures in detail; however, some important metrics are listed as follows that are used in this review to evaluate the performance of models.

Accuracy is the most heavily used measures to evaluate any machine learning model. In case of classification problem, accuracy is the ratio of correct prediction to the total number of examples. For a two class classification or binary classification problem, it can be represented using Eq. 13. True positive, True negative, False positive, and False negative are abbreviated as TP, TN, FP, and FN respectively. However, mean squared error (MSE), root mean square

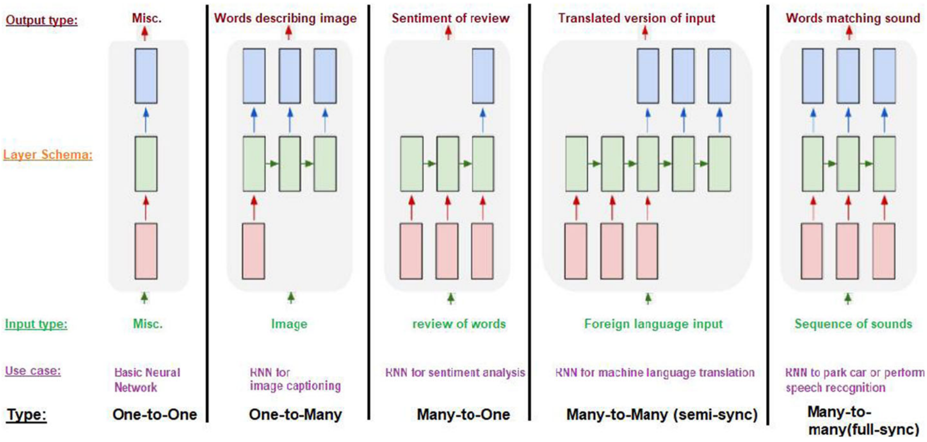

Fig. 6 Types of RNNs [127]

error (RMSE), and mean absolute error (MAE) are some of the important performance measures in case of regression problem.

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

Precision is second most important performance measure in this field to evaluate the model focuses on false positives (FP). For a binary classification problem, it indicates how many samples are truly positive among all positively predicted samples, and can be represented with the help of Eq. 14.

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

Recall is another important performance measure, generally used for the evaluation of the models focuses on false negative (FN). It indicates that how many samples are truly positive among all the actual positive samples. This measure can be represented using Eq. 15.

$$Recall = \frac{TP}{(TP + FN)} \tag{15}$$

Ideally, FN and FP should be zero in case of binary classification problem. This can be extended in the similar fashion for multiclass classification.

F1-score is also an important criterion to evaluate the model, taking into consideration of both the previous two performance measures i.e. precision and recall, and can be described as the harmonic mean of these two measures, and expressed in Eq. 16.

$$F1\text{−}score = 2\left(\frac{Precision*Recall}{Precision + Recall}\right) \tag{16}$$

4 Sequence models in textual sentiment analysis

The automatic or manual extraction and analysis of the subjective information such as sentiments, emotions, or attitudes from textual input using NLP techniques is collectively called as textual SA. The input to this process maybe in the form of characters, words, phrases, sentences, paragraphs, documents, or any combination of these in any language. Herein, we cover the issues and challenges, textual input representation, word embedding, contextual embedding, and the utility of RNN in textual SA.

4.1 Issues & challenges

Extracting the associated sentiments from the textual input is not simple, as it depends on and is determined by several critical factors. The various issues and challenges in textual SA determination are categorized as follows:

Subjectivity information: Capturing the subjectivity, determining opinionated information if any, and extracting the intensity of the associated emotion from textual input is an incremental NLP task. In general, but not necessarily, subjective sentences have an opinion or emotion associated with it. Table 4 offers an understanding of sentence or phrase subjectivity and its deterministic aspects in the context of opinions and polarity sketching.

Level of input: Broadly, the input to textual SA is categorized into three types; document-level, sentence-level, and phrase-level. In Table 4, all examples are either sentence-level or phrase-level. In document-level SA, different opinions in different sentences are possible for the same entity. Consider the following document-level review where different opinions are given for the same entity, reflected in different sentences or phrases of that document:

S1: I am very happy today because of purchasing new iPhone. I like this iPhone. My sister doesn't.

Moreover, aspect-level SA specifies the sentiment or emotion for the same entity with different aspects. The following sentence contains different aspects of the same entity and it is difficult to extract the cumulative sentiment from such sentence.

S1: I like the thickness of this iPhone but require more power.

Contextual information: Performing SA in the right context is also a challenging task. Context plays an important role in mining the correct sentiment. For example, India may refer to a country, the largest democracy, or simply a cricket team. Another example is the following sentence which might be an assertion of any one of the *n* sentences that could exist:

S1: It is absolutely amazing.

Short sentences such as tweets lack contextual information and it is difficult to extract the exact semantic of the sentence. These short texts can be on any topic, part of a conversation in a group, or a comment on any subject that has a different context. Several consecutive statements in a conversation and a set of specific statements on any topic are very useful for capturing conversation-based context and topic-based context respectively. Conversation-based context is quite clear whereas the same hashtags in tweets may be considered as an example of topic-based context. The third category is the author-based context that contains statements belonging to the same author, basically useful in predicting author characteristics such as personality.

Hidden irony: Sarcasm, i.e. hidden irony in the text may reverse the associated polarity. It is the most difficult part in SA and lots of models and approaches are presented in the literature to detect sarcasm. Consider the following example that actually has a negative sentiment, but the overall sense seems to be positive:

S1: The features of this iPhone are too good to handle.

Table 4 Subjectivity Information

Sentence	Subjectivity	Subjectivity Intensity	Opinion	Polarity
I bought an iPhone yesterday	×	*NA*	×	*NA*
My iPhone got damaged	×	*Low*	✓	–
My iPhone is good	✓	*Low*	✓	+
I think my iPhone shall be delivered tomorrow	✓	*Low*	×	*NA*
My iPhone is so nice	✓	*High*	✓	+

4.2 Input representations

Effective and meaningful representation of textual words has its own importance in text data processing. This effectiveness is propagated into the phrase-, sentence-, paragraph-, and document-representations. Broadly, word representation is classified on two bases namely *frequency-based* and *prediction-based*. Count vector, TF-IDF [9], and co-occurrence vectors [20] were quite popular in frequency-based methods. Each of these methods is solely based on the count of the words in a text segment. Prediction-based word representation has been hugely popular due to the lack of predictions, semantic knowledge, and contextual information in frequency-based word representation. In the prediction-based approaches, the widely adopted technique for representing text data is embedding (distributed vectors) i.e. a dense vector representation of the text. On the other hand, sparse vector representations such as *one-hot encoding* is computationally expensive, less effective, and disable to capture similarities among words. Generally, embedding dimension is very less as compared to vocabulary size as it is completely based on the Featurized representation. Elements of a dense word vector signify the weightages corresponding to that feature. A character, a word, a phrase as well as a sentence can be encoded using a distributed vector. Among all embedding techniques, word embedding is quite popular. The basic idea is that similar or related words such as *plane* and *aircraft*, *king* and *queen*, and *male* and *female*, must have similar real valued vectors i.e. cosine similarity of these vectors pairs actually illustrate the semantic and morphological similarities. Each word is represented by a real valued vector and a phrase/sentence is a concatenation of several words. Character-level embedding is the very basic form of embedding that represents a vector for each character in the alphabet, captures the morphological information, and avoids the *out-of-vocabulary* problem that may be the case in word-level embedding. Lesser parameters are to be learned in character level embedding as compared to word-level embedding. Conclusively, text embedding is an NLP technique basically employed in feature learning and language modeling where textual segments are mapped to real valued vectors that play a major role in textual SA.

Word2Vec [101, 116] is the most popular distributed representation of the words over a pre-specified dimensionality space that can be implemented using a shallow neural network. It takes a large corpus of words as input and produces vectors for each word such that similar words are in close proximity. The efficiency of Word2Vec can also be measured in a way such that the word vectors of *king* and *female* produce the word vector of *queen*. There are two models that act as a part of Word2Vec algorithm, namely continuous *bag-of-words* (CBoW) and *skip-gram* model. These two models respectively predict one target word from the surrounding context words and multiple context words from the given target word as sketched in Fig. 7. The importance of context words is very much clear from the famous quote by J. R. Firth "*You shall know a word by the company it keeps*." The choice of the model for the implementation of Word2Vec depends on several factors such as the size of dataset, word property, and data property used in embedding training. Generally, CBoW and skip-gram models are well suited for smaller and larger datasets respectively. It is mandatory to define a vocabulary of known words while implementing any of these embedding models. The efficiency of the embedding depends on the size of the training data. As the training data may not always be so large in a SA task, the learned word vectors might not represent good embeddings. A better solution to avoid this problem is to load pre-trained word embeddings constructed on a much larger dataset. Unsupervised approach for learning word embeddings is an impressive and largely adopted technique used for word representations.

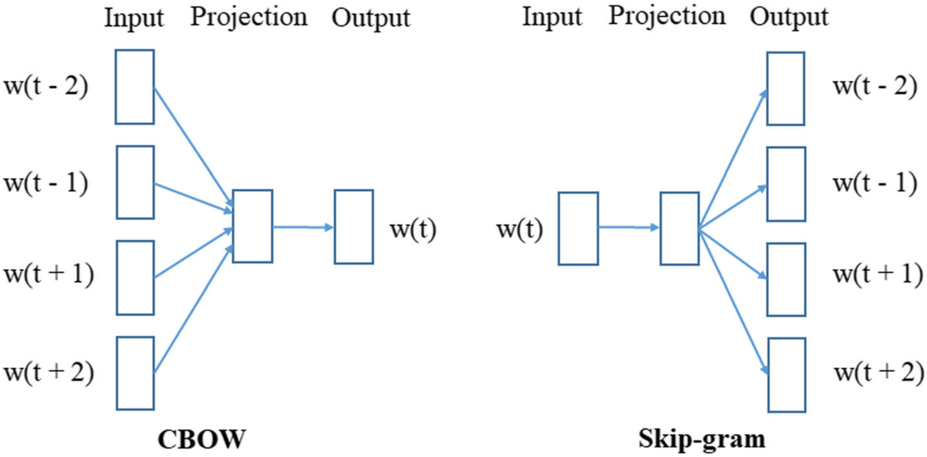

Fig. 7 CBoW and Skip-gram model

The context choice (forward or backward or bidirectional) and context size define the parametric complexity of the shallow neural network in Word2Vec implementation. Initially, the input and output are represented using one-hot encoding, a vector of vocabulary size. In the shallow neural network, the number of hidden neurons is quite less than the vocabulary size. In this way, the output of the hidden neurons is actually called word embedding, which is much compact and less sparse in space while maintaining enough of the original information. Similarly, for a specific context size, same number of word embeddings are generated which may be averaged or max pooled. In some cases, unity context size may not necessary get the actual semantic such as in the phrase "*cool dude*".

Although, character-level embedding and word-level embedding are quite popular, context-level embedding [108] is mostly adopted in SA. The reason is that the same word may signify different contexts in different sentences. Different embeddings for that word should be produced in each sentence specifying the right context. Consider the following three sentences with same word *date* having different senses in each sentence.

S1: His favorite dry fruit is date.
S2: The date for the election is very close.
S3: He is dating her first time.

4.3 RNNs in textual SA

The RNNs witness the overwhelming success in determining and improving accuracy in textual SA in the recent past. In this section, we highlight several interesting aspects in SA using RNNs. More specifically, we elaborate the uses and applicability of RNNs and its architectural variants in the textual SA. We broadly classify this section in the following two categories:

4.3.1 Architectural variants and hybridization

The architectural variants of RNNs are very popular for enhancing the SA precision. The modified RNN with dual feedforward networks is presented in [117] that takes input as a segment and stores long history across the time in the memory. In addition, all the statements provided as input were considered to identify the overall polarity. In [121], a bi-directional LSTM-RNN is applied to perform robust segmentation and classification jointly. The efficient computation of segmentation using RNNs is achieved which is not affected by the presence of *linguistic features*, *sentence boundaries*, and *punctuations*. After the segmentation, opinions or corresponding sentiments are extracted from the text data.

In deep neural networks, connections play a vital role to determine the flow of information. These connections substantially reduce the effect of gradient problems and enhance the capability of model learning. A new delay connection without any extra parameters is introduced in a LSTM called as Delay Connected LSTM (DCLSTM) [135]. The DCLSTM maintains the output of a LSTM, this functionality lacks in a LSTM. A DCLSTM also, leverages to handle error signals to previous steps, back propagated to different layers without vanishing rapidly. The DCLSTM model is shown in Fig. 8 consist of the three inputs: x – external input, y – output of recurrent unit, and s – memory state. Moreover, input gate (i), output gate (o), forget gate (f), and hyperbolic tangent function (g) are semi-linear units are adapted respectively. The delay block with respective element wise operations such as addition (+), minus (1-), multiplication ( ) with a hyperbolic tangent function (*tanh*) are also indicated in DCLSTM. LSTM with new quadratic connections is presented in [140]. This LSTM model can be utilized for SA and semantic relatedness. Input to non-leaf nodes comprises; outputs from left children ($c_{t-1,\ l}$), right children ($c_{t-1,\ r}$), and two forget gates ($f_{t,l}, f_{t,r}$). The inputs and forget gates ($f_{t,l}, f_{t,r}$) are the composition of hidden vector as well as quadratic terms jointly by non-leaf nodes children's as shown in Fig. 9. The linear connections in standard LSTM are unable to capture the complex semantics of the given text. These semantics lies between words i.e. sentiment strengths or negated sentiment. In [50], a mixed model approach is presented to identify the sentiments from sentences. Here, to overcome the problem of CNNs i.e. to stack multiple convolutional layers for capturing long-term dependencies, a joint CNN and RNN framework is proposed by the authors that uses word embeddings as input. The outstanding results are achieved on SA benchmarks with hyperparameter tuning and static vectors by mixed model. Another integrated model of CNN and LSTM is developed by [69] for analysing the posts on social media to predict real-time sentiment. The opinions and facts are separated automatically by the proposed system wherein single layered CNN utilized for convolution operation and two-layered LSTM is applied for raw data representation. The accuracy, precision, recall, and F-measure are reported as 91.82%, 86.21%, 91.52%, and 88.20% respectively.

Recently, Tree-based LSTM is proposed in [78] for SA to overcome the problem in previous works where sentence structure and respective words carries dependencies among them were not researched intelligently. Mostly, models are not able to distinguish the change in meaning when semantics of sentence is altered and called as typed dependencies which are related to sentence structure. Relation gated LSTMs (R-LSTMs) are utilized to learn semantic of sentence with dependencies, also used to regulate the hidden state in LSTM. In [48], SA in big data environment is proposed by using RNNs for fastText. The objective of this study was to improve the performance of RNNs in automated distributed environment for big data related to social platforms. The authors show that how we can store, manage, and visualize the data in real-time coming from different sources where learning takes place in distributed environment. Moreover, LSTM, Bi-LSTM, and GRU are modified to do the task of text representation and sentiments classification.

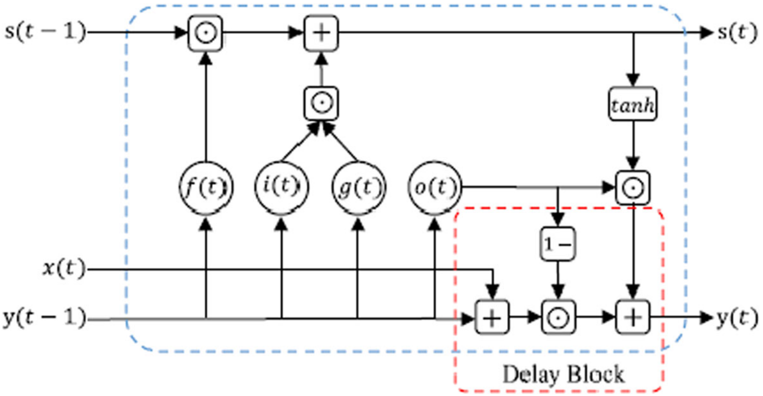

Fig. 8 DCLSTM architecture [135]

4.3.2 Application level

Due to the huge popularity of **Twitter** platform all over the world, notable results are drawn for different languages of Twitter datasets in textual SA. The literature also traces that there is no language barrier for the extraction of sentiments or opinions. SA for Chinese Tweets [23], Japanese [105], Spanish Tweets [4], Arabic [2] have been investigated by various researchers using RNNs. A global RNN [23] is proposed for Sentiment Classification task on Chinese text in which outputs of all the timesteps are utilized as features to extract the contextual information. The aspect-based sentiment prediction comprises two subtasks viz. aspect terms extraction and identification of sentimental polarity of those extracted aspect terms. Aspect

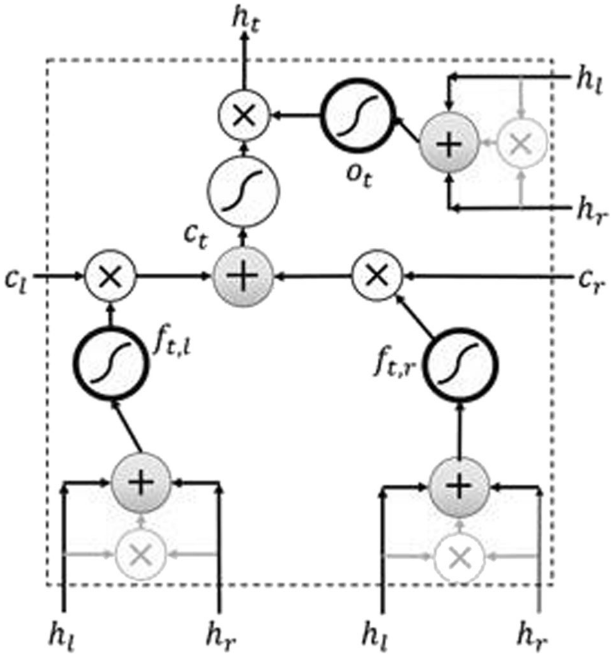

Fig. 9 Architecture of LSTM cell with quadratic connections [140]

terms for Russian and English datasets are extracted using bidirectional LSTM [130] that outperforms Conditional Random Fields (CRF).

Eventually, **Microblogs** attracted the people to express their views on different topics of interest such as *services*, *products*, *personalities*, etc. RNN is employed to identify the Chinese public figure using SA [24]. The traditional SA methods for microblogs were not able to predict the opinion polls for public figures in question. However, parsing-based SA architecture [24] is proposed that jointly use the targeted opinions and their related sentiments to overcome the mismatching between them. The model comprises three different steps; 1) *data collection*, 2) *parsing sentiment*, and 3) *aggregating opinions*, wherein RNN is used for sequence labeling task for the input, as illustrated in the Fig. 10. Context attention-based LSTM (CA-LSTM) [41] is developed to process the microblogs as a sequence; a hierarchical structure to handle microblogs and attention mechanism for words and by providing different weightages to tweets, in SA process. Further, a short text SA for word vocabulary using word sensitive LSTM is presented in [58]. The underlying keywords influence the semantics of a given document. The modified LSTM and GRU are used in enhancing the memory of keywords. Additional information i.e. *keywords* are passed to the input gate and forget gate change those inputs. The performance is evaluated on SemEval-2016 and IMDB and the proposed model outperforms the basic LSTM.

The SA is not only be applying to tweets as a parameter but can also be applying to different dimensions such as volume, sentiment, and influence of tweets. A three-dimension information diffusion model for SA is demonstrated that recognizes the patterns and modeled to quantitatively predict on Twitter datasets [51]. The time series clustering is utilized to discover various patterns and LSTM is used as a prediction model. The results obtained using LSTM are better as compared with ARIMA [54].

Furthermore, deep learning is also used for analyzing the people **health** issues. The personal experience tweets are very important for health surveillance. Well annotated balance dataset is the first and foremost requirement for such mining, proven to be a labor intensive task. Authors present a filter based machine learning approach capable of producing such tweets and maintaining balance among the classes with a reduced annotation work [65]. A new approach is investigated using combination of word embedding and LSTM to identify SA of medicine information and consumption based tweets [66]. The word embedding is used for

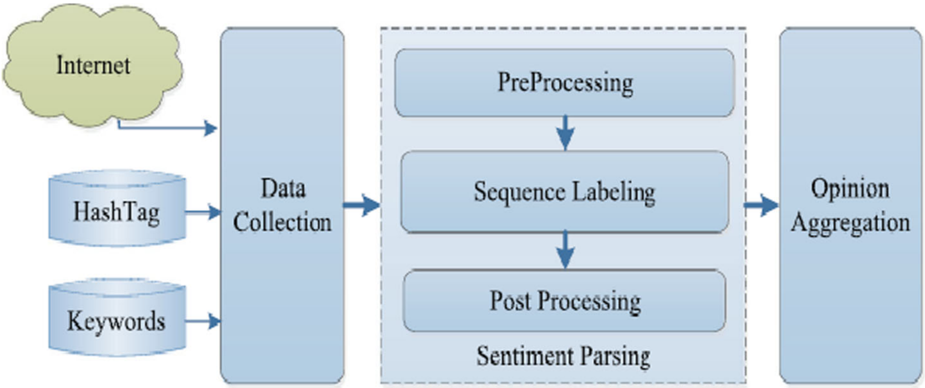

Fig. 10 Parsing-based SA architecture for microblogs [24]

dense vector representations using vector space model as an index for tweets and LSTM is employed for sequence processing and classification.

The correlation between **stock market** and sentiment has been studied in [89]. The SA using RNN for the prediction of stock volatility and factors affecting the stock volatility such as sports, weather, election, political drama, etc. are discussed. For each online financial post related to specific stock, a sentimental score is computed by the proposed model. The accuracy of the SA prediction is improved using deep learning methods for Google Play consumer reviews [30]. The LSTM cell is used for the SA process whereas LSTM is utilized to train the prediction model. The model achieves an accuracy of 94% that outperforms Naïve Base (NB) (74.12%) and SVM (76.46%). Nevertheless, predicting the stock price correctly on given day is a challenging task. The stock price prediction based on wavelet transform and LSTM with attention mechanism is presented in [112]. The wavelet transform is used for denoising the historical data and perform the data normalization. Moreover, LSTM-attention applied to predict the final sentiments.

A personalized framework for **individual-centric** SA uses the individual's past history is proposed in [47]. All individuals consist some set of uniqueness which is the basis of this framework. Each individual have different lexical choices i.e. an indicator for individual for predicting relations between the documents and user sensitivity. The RNN is applied to learn the individual's past sentiments, to determine dependence of individual's topic, and to identify hidden relations amongst individuals. The deeper understanding of the **behavioral** aspects of an individual is modeled in predicting SA.

To mimic the human thinking, reinforcement learning is adapted by [22] on sentences where sentiment inclination is determined amongst the words. The different LSTMs are appointed for the identification of activities such as *neutral*, *positive*, and *negative*. Moving over these LSTMs, sentence-level representation and sequence of sentiment in word-level is obtained which is used for sentiment classification. In [107], sentence-level sentiment prediction is demonstrated which targets the challenges of finding information from text associated with subjectivity. The two models are used to do the comparative analysis i.e. LSTM with dense layer and deep feedforward NN with pooling.

The summarization of SA using RNNs based on *SA Tasks*, *key characteristics*, and *dataset* used, is presented in Table 5.

5 Sequence models in visual sentiment analysis

Visual SA refers to the extraction and analysis of the associated polarity or emotions depicted in the visual inputs. Visual inputs may be images or videos, posted on any social media. The sentiment polarity and fine-grained classification with the help of visual features of images or videos may sometimes overcome the linguistic barrier. In this section, we present the challenges and applicability of RNNs in visual SA.

5.1 Challenges

Categorization of the visual input is characterized on the basis of the count, size, and strength of various features. The visual SA determination challenges can be broadly classified into four categories: 1) features identification and extraction, 2) heavy computation for the processing of high-resolution data, 3) Integration of the processed features, and 4) extraction of the polarity and associated emotions from these processed features.

Table 5 Summarization of textual SA using RNNs

Ref.	Compared with	Applications	Key characteristics/contributions	Datasets & Results
[137]	Bi-LSTM, Deep Bi-LSTM, Residual Bi-LSTM	Sentiment Classification	• Combination of word embedding and residual Bi-LSTM • Deeper RNN (Bi-LSTM) architecture	• CTAS: 9178 negative and 2363 positive samples, accuracy - 92.12% • JTCC: 86061 positive and 9790 negative samples, accuracy - 98.23%
[167]	FastText, TextRNN, TextCNN	Text Classification	• Framework for document representation based on hierarchical architecture • Hierarchical neural models for document classification i.e. TextHFT, TextHRNN, and TextHCNN	• Yelp 2016, accuracy: TextHFT – 59.27%, TextHRNN – 59.88%, TextHCNN – 62.15% • Amazon Reviews (Electronics), accuracy: TextHFT – 61.05%, TextHRNN – 59.78%, TextHCNN – 60.32%
[118]	[28, 67, 68, 102, 164]	SA	• Bi-LSTM for text classification based on supervised and semi-supervised approaches • Combination of adversarial, virtual adversarial losses and entropy minimization loss for unlabeled and labeled data	• ACL-IMDB, 26.9% reduction in error, accuracy – 95.53% • Elec • AG-News topic, 26.6% reduction error • DBpedia
[41]	RNN, LSTM, Bi-LSTM, GRU	SA	• Hierarchical structure for modeling microblog sequences • Different weights for the words and tweets by using attention mechanism.	• COAE-2015, accuracy - 65.18%
[50]	BoW, SVM-bi, NB, SVM, RNTN	SA	• Combination of CNN and RNN framework • Feature maps learned through LSTM	• IMDB, accuracy – 93.20% • Sentiment Treebank (SSTb), accuracy – 89.20%
ACNN [91]	NBSVM, Paragraph-Vec, LSTM with tuning and dropout, BRCNN	Sentence representation and classification	• Combination of bi-attention and CNN • Forward and backward RNN with attention to learn forward and backward context vector	• IMDB, accuracy – 91.10% • TREC dataset, accuracy – 95.50%
Res-RNN [153]	RNN, LSTM, GRU	Learning Sequential Representations	• Novel recurrent unit with residual error • Residual learning is used to solve the gradient in LSTM	• ATIS database, accuracy – 96% • IMDB database, accuracy – 95.12% • Polyphonic music database, accuracy – 94.23%
TD-biGRU [63]	SVM-dep, LSTM, TD-LSTM	Target-dependent SA	• Identifying and extracting the target from the tweet • Identifying the polarities of the tweet based on each extracted target	• [38], accuracy – 72.25%
[82]	GloVe, CBOW with hierarchical Softmax (CBOWHS),	Word Embeddings for SA and Sequence Labeling	• ELM-based Word Embeddings for text Categorization i.e. SA and Sequence Labeling • Count based method similar to GloVe model	• Sentiment Polarity Dataset v2.0: 1000 positive and negative reviews • ATIS dataset

Table 5 (continued)

Ref.	Compared with	Applications	Key characteristics/contributions	Datasets & Results
	CBOW with negative sampling (CBOWNS), Skip-gram with hierarchical softmax (SGHS), skip-gram with negative sampling (SGNS)			
[66]	KNN, SVM, Decision tree	Efficient Word Embeddings for SA	• Combination of word embedding and LSTM • Word embedding used for dense vector representation • LSTM used for sequence processing and classifier	• Corpus of annotated tweets [28], accuracy – 81.50%
[134]	RNN-Cap	SA	• The model consist of three modules: representation, probability, and reconstruction • RNN is used as hidden vectors encoding	• Movie Review (MR), accuracy – 83.80% • Stanford Sentiment Treebank (SST), accuracy – 49.30% • Hospital Feedback dataset, accuracy −91.60%
WSLSTM [22]	CNN, LSTM, RCNN, ID-LSTM	Word-level sentiment analysis	• Reinforcement Learning	• Movie review: 5331 positive and negative reviews, accuracy −78.9% • Amazon Food: positive reviews - 5 point grade and negative reviews - 1 point grade, accuracy – 93.5% • Amazon Mobile: Similar to AF, accuracy −95.1% • Word Vectors: nlp.stanford.edu
ICNN-LSTM-DNN [69]	CNN, LSTM, Logistic Regression	SA	• Integration of CNN and LSTM model • One layer of CNN • Two layers of LSTM	• SNAP • 100 K • T4SA • Sentiment 140 • CKAN dataset • Accuracy - 91.82%
Tree-LSTM [78]	LSTM, BI-LSTM, GRU, DT-RNN, DT-LSTM, DT-GRU, D-LSTM, SDT-RNN	Sentence Semantic	• Capture relation type among the words in input sequence • Enhanced representation of semantic of sentences based on dependency between sentence structure and underlying words	• Stanford Sentiment Treebank (SST), accuracy - 86.4% • SICK
[48]	LSTM, Bi-LSTM, GRU XGBoost-avg-fastText	SA	• fastText representations to accomplish text classification • Distributed learning using LSTM • Bi-LSTM and GRU for sentiment classification	• Yelp, accuracy - 78.88% • Sentiment140, accuracy - 93.28%

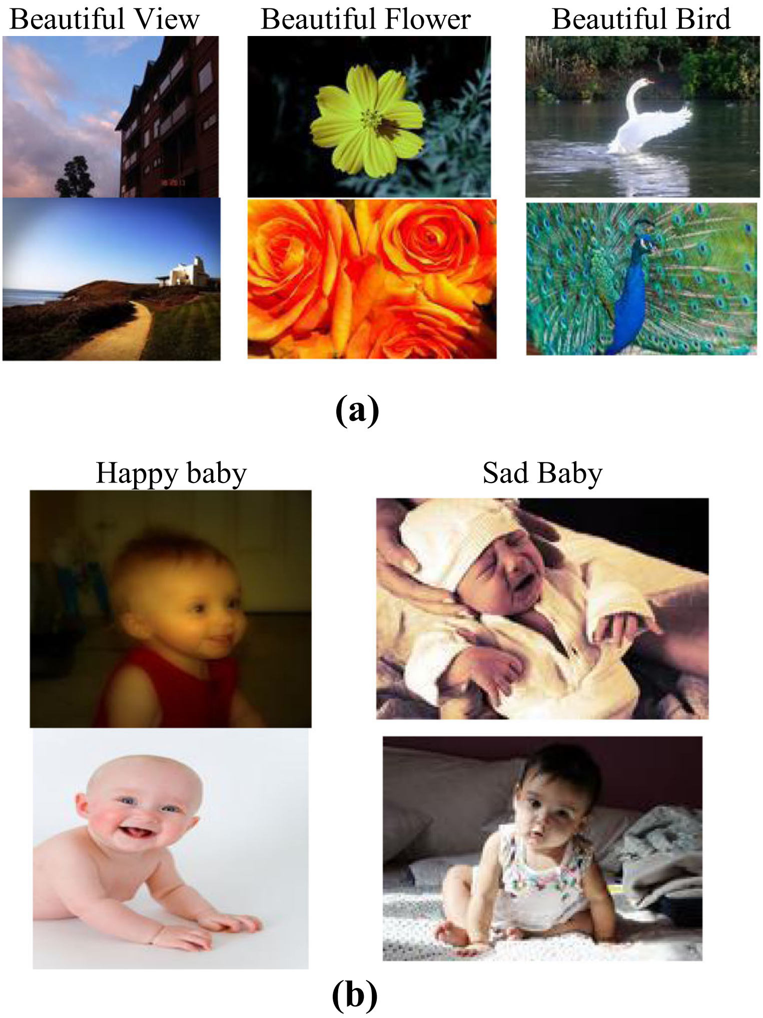

Fig. 11 **a** Intra-class variance, **b** Object with different sentiments

Intra-class variance is very large in visual SA determination as positive or negative emotions may map to millions of objects as depicted in Fig. 11a. *View*, *flower*, and *bird* are three very different objects, but all belong to the same positive sentiment. Emotion or sentiment corresponds to high-level abstractions and subjectivity for a visual input [8]. These high-level abstractions may also demand additional knowledge in viewers. In other words, visual SA determination is another layer superimposed on the object recognition task. However, with little change in the image, sentiment polarity and the associated emotion may be completely reversed in visual SA, which is not the case in object recognition. As shown in Fig. 11b, the object is same in each image, i.e., *baby*, but they have opposite polarity of sentiment. Sometimes, visual SA may be considered simpler as compared to image classification only in the context of the number of output categories, i.e. two to three in a coarse-grained SA and approximately five to eight in a fine-grained SA. However, according to [37], the categorical outputs for ED are quite large, i.e. 24 emotions.

Another aspect is the human thinking ability over the visual input. Two persons may have totally different sentiments regarding the same visual input. These become worse in acquisition of labeling the visual entity. Mislabeled input is one of the major sources of noisy inputs in visual data processing. Other sources of noisy inputs in visual SA may include deteriorated images and annotations discrepancies. In some images, polarity and emotion is reflected by a complete spatiality of the image such as that of a *pleasant environment*. However, in some images, polarity and emotion are completely dominated or determined by only a little spatial part of an image such as that of a *barking dog*.

5.2 RNNs in visual SA

Visual SA is gaining importance due to the rapid increase in visual contents on social media. The prediction of SA on images is challenging task and is discussed in [21]. Representing information using visual aids plays a vital role in analysis and decision making. In [60], different visualization techniques are used for SA to generate the sentiments or opinions from Twitter dataset of *Government-Citizen* interactions. The similar study is conducted in [92] on visual analysis of geo-located Twitter data for sentiment visualization. This model is a combination of SA and geographic visualization.

To deal with visual pattern, RNN was applied as a predictive model for generating and recognizing dynamic visual patterns under predictive coding framework with principle of error minimization [26]. The predictive multiple spatio-temporal scales RNN (P-MSTRNN) is an extension of [129], wherein no prediction mechanism was present but with P-MSTRNN, we can learn, generate, and recognize the patterns. Moreover, P-MSTRNN consists of context layers utilizing CNN and Deep CNN for feature extraction. Sigurdsson et al. [124] presents an automatic identification of sentiments from visuals inputs (images) using RNN. Skipping RNN (S-RNN) framework does not predict every data point in the given sequence as in case of simple RNN. An efficient sampling method is adopted to discover the accurate storyline i.e. skip through the images. S-RNN outperforms LSTM for its learning capability of long-term correlations and recognition of latent storylines. In addition, high correlation problem between consecutive images or photos is overcome by S-RNN.

The fundamental challenge in visual SA is the recognition and identification of contents. Sentiments are predicted using Long-term Recurrent Convolutional Network (LRCN) model [36] wherein visual features are extracted from visual input through CNN and layers of LSTMs are applied for sequence learning. The LRCN can be applied to various tasks which are shown in Fig. 12. Furthermore, various methods only focus on visual information present in the video in SA. To generate accurate descriptive sentence, we need to incorporate the audio cues in synchronization. The impact of short-term and long-term dependencies is utilized and visual-audio mix-modality model employ the extended RNN with internal memory to preserve the important information [49]. The sentiment i.e. positive, negative, or neutral is discovered from an image, the methodology for SA is presented in [132]. The sentiments from the text are utilized for classification of visual sentiments. A LSTM-SVM based model is adopted; LSTM used for features extraction and storing temporal dependencies whereas SVM acts as a classifier.

Automatically generating sentiments from an image has two challenges i.e. *language* and *vision*. The model is proposed [97] to generate captions with sentiments i.e. positive or negative from a given image. A switching RNN model is presented with word-level regularizer wherein a combination of two parallel CNN-RNNs is used. One is used for accurate word generation and other for sentiment classification. Here, RNN consist of series of LSTM cells and switching takes place between RNNs.

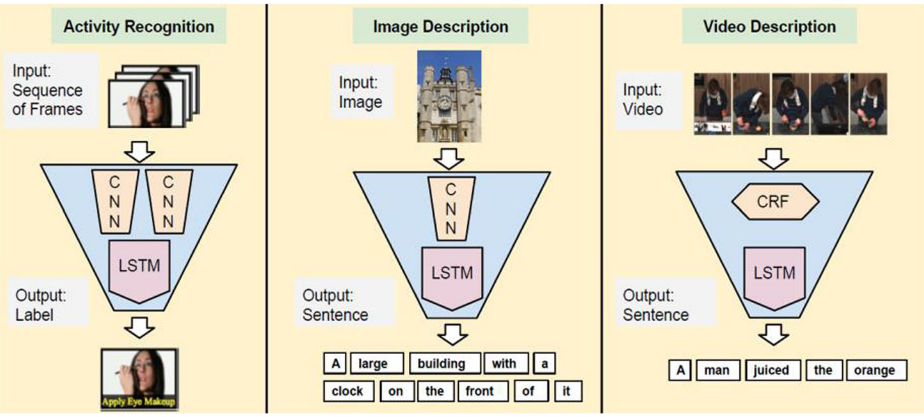

Fig. 12 LRCN Model [36]

An aspect based SA is the fundamental task that predicts the sentiment polarity based on given aspect. The proposal for aspect based SA for images and text is discussed in [146] wherein information indicated in image is suggestive to text. Multiple correlations are captured from the different modality data based on; 1) aspect-level SA, 2) aspect impact on text and image, and 3) relation between text and image. Multi-Interactive Memory Network (MIMN) consists of Bi-LSTM which is used for aspect feature embedding and textual and visual memory building. The textual and visual attentions are used synchronously from respective memory model i.e. textual and visual memory model, for learning efficient interaction.

Nevertheless, it is found that the use of *unicode* for expressing emotions related to sentence is increased and standardized. The emojis utilization is common that are best way to express emotions in sentences. A new model is proposed [131] that identify the emojis as labels for the given sentences. GRU as recurrent network is utilized as an encoder for sequences and decoder for word embedding. Moreover, encoder-decoder network works as a classifier for predicting the correct emojis. Furthermore, recognition of emotion from image using features fusion is demonstrated in [168]. The emotion in image can be affected by various attributes such as object texture and color, etc. These features taken independently at different level with their dependencies and then features are fused. The CNN and RNN based model [168] is utilized for emotion detection wherein CNN is utilized for feature extraction at different levels and features fusion is performed by Bi-RNN. Result reported on datasets Art photo and Internet Image using CNN and Bi-RNN highlights 7% increase in performance compared to similar work. In [85], hierarchical combination of CNN and RNN is analyzed which is the extension of previous work by [168]. Here, CNN is applied to learn features from various levels i.e. local to global and the stack of Bi-RNN are utilized to aggregate the features at different level by discovering the dependency of features at various levels. After experimentation, an improvement of 13.2% is achieved over [168] on different datasets.

6 Sequence models in multimodal sentiment analysis

The researchers applied the RNNs not only for textual or visual SA but for the combination of text, video and audio modality i.e. multimodality. The some of the notable research is highlighted in this section for multimodal SA.

6.1 Challenges

Multimodal SA is the extraction of associated sentiments and emotions from the multimodal inputs. Nowadays, most of the posts or blogs on social media are multimodal i.e. more than one modality is associated with the data, mostly the combination of textual and visual. Though, training of the deep neural network for multimodal SA is difficult, accuracy is enhanced considerably on the other hand, if the model is trained properly. In the previous two subsections, we explored the techniques for unimodal SA i.e. for *textual* and *visual* input separately. The challenges in the feature extraction and processing of multimodal SA can be categorized as follows:

A Processing of joint modality is the first challenge in multimodal SA i.e. how should we separate out the modalities? In addition, multimodality seems to be computationally expensive as we have to separately extract and process the features for each modality.

B Features extraction for multimodal SA also plays an important role. In other words, each modality is processed individually. The next question is that up to what extent they are processed individually? The first option is that extracted features are combined and these combined features are provided as input to the classifier. It is much more challenging to merge the features having different modalities and these different modalities should be consistent in terms of depicting the same subject. Second option is extracted features are provided to different SA engines/classifiers depending on the modality and local sentiment scores/decisions are combined to form a global decision. These two methods are known as *feature level fusion* and *decision level fusion* respectively, as illustrated in the Figs. 13a and b respectively.

C Assigning weightages to each modality is another important point to be considered carefully. In *feature level fusion*, feature vectors of highest weightage modality will be a dominating entity in the fusion process. Sentiment score of the highest weightage modality should also be taken care in *decision level fusion* while clubbing different sentiment scores.

6.2 RNNs in multimodal SA

An **emotional** state corresponding to the user is identified rather than a *positive* or *negative* sentiment from the sentence description [56]. The images are handled by *inception* model and LSTM is used for word embedded in text processing. These outputs are directed to the dense layer for the generation of more appropriate sensible list of words by which emotions are detected automatically.

In [6], different feature learning approaches/models based on neural networks such as skip gram and denoising autoencoder are investigated for the Twitter Multimodal Dataset in SA determination. An extension of CBOW is proposed that learns textual features using concurrent vector representation and visual features with the help of denoising autoencoder. Machine learning approaches in combination with language based formalization is proposed for extracting sentiment polarity from multimodal data wherein formalization of data for the experimentation is performed using multimodal language [17].

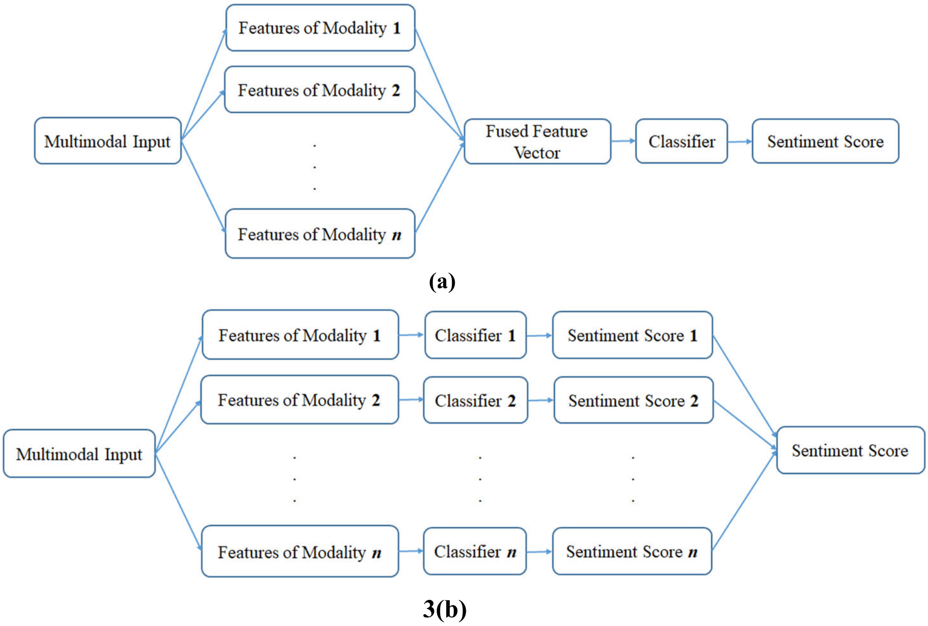

Fig. 13 **a** Feature level fusion, **b** Decision level fusion

Feature vectors from textual, visual, and speech modality are extracted and combined to represent the effective multimodal **features** [110] i.e. authors select video clips of speaking people for this experiment. In addition, a comparative study between the two aforementioned fusion approaches is presented. In feature level fusion, authors add modality in the incremental fashion, and accuracy is marked for each combination of modalities. Both of these approaches outperform to the state-of-the-arts. Authors also conclude that selection of important i.e. dominating features from each modality is very important task in multimodal SA.

Liu et al. [90] developed a multimodal fusion method wherein a low-rank **tensor** is adopted instead of regular tensor due to exponential growth in dimensions and high compute complexity while transforming input into tensor. Another tensors flow network in [154] demonstrates the use of dynamics i.e. intra-modality and inter-modality for end-to-end fusion which analyze sentiments represented in unimodal, bimodal, and trimodal forms. Here, LSTM is utilized for learning language representation over the time from GloVe [71] word vectors.

Pre-trained CNN model on Imagenet [150] is employed to fine-tune the CNN for visual portion of the multimodal input [32]. Textual features associated with each image are learned through distributed representational word embedding. Multi-modality regression model is employed to impose the consistency among all the modalities. In this series, logistic regression is employed to fuse the probabilistic results of textual and visual modality [151]. Sentiment prediction on text are having more accuracy as compared to visual SA, therefore, more weightage is assigned to textual sentimental score and textual features in *decision-level* and ***feature****-level fusion* respectively.

To date, SVM was the only model in single kernel category for fusion of different features i.e. modalities. Multiple Kernel Learning (MKL) is a feature selection method where similar features are grouped and each group has its own kernel. Multiple kernels are employed for fusing the audio, video, and textual modalities [111] on YouTube videos. Temporal CNN is employed for capturing the video features in which visual input is considered in the form of five to ten seconds utterances. Each utterance is annotated with a coarse-grained polarity.

The more complex multimodal analysis is face-to-face communication. Human can easily understand the words, gestures, and tone in face-to-face communication to comprehend the sentiment of each modality. The neural network based approach called Multi-Attention Recurrent Network (MARN) proposed in [156] that understands communication in all modality and generates the corresponding true sentiments. The LSTM is used for each modality as a memory i.e. each modality will store view-specific dynamics and cross-view dynamics. Subsequently, Seq2Seq Modality and Hierarchical Seq2Seq Modality translational models for multimedia data are presented in [109] which are used for multimodal representation learning. The RNN/LSTM is utilized in various phases as modality encoder, modality decoder, and sentiment predictor. Moreover, the multimodal SA is further utilized for the task of personality detection from text [94] and identification of regions of interest of audio signals from songs data [1].

Nevertheless, the nonverbal behaviors, visual and acoustic patterns are analyzed and sentiments are recognized by using nonverbal sub-word and dynamic word representation [133]. The separate LSTM is adopted to compute visual and acoustic embedding corresponding to word embedding. For sub-word sequences, modality-specific LSTMs are used. LSTM is also applied as a sentiment predictor in the last phase when multimodal word representation is done for each modality. This fine-tuned model behaves like having the human understanding. Majumder at el. [95] applied the GRUs for multimodal SA in which unimodal features are fused in hierarchical manner. The GRUs are utilized at the starting layer of hierarchical model for the generation of context-aware features for video, audio, and text respectively. Then textual features are extracted using CNN from each utterance, openSMILE [39] is adapted to extract audio features, and 3D-CNN [64] is applied to extract the features from videos. The performance is evaluated on Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) [157] and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [16] along with an accuracy of 80.0% and 76.5% respectively. Another, use of RNN for multimodal SA is highlighted where Bi-GRU multimodal attention platform is presented in [42]. Here, contextual information plays a vital role in multimodal SA related to text, audio, and video. Subsequently, attention mechanism is applied on multimodal data to capture underline features amongst them. The benchmarking results are obtained on CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) and CMU-MOSI with respective accuracy of 79.80% and 82.31%.

Recently, multimodal SA is proposed by [142] wherein complex correlation between text and image has been investigated. The semantic and visual attention is adapted to identify words and regions of emotion from *image-text* pair. The multi-level interaction between semantic attention and visual attention give the visual features and vice versa. Further, these features jointly predict the correct sentiments from multimodal input. Moreover, CNNs is applied to find region maps and LSTM is used as an encoder for sequence of texts. Similarly, semantic and visual contents correlation is utilized to extract the promising features from image-text pair using attention fusion in multimodal data [59]. Two independent unimodals with attention mechanism are applied to acquire effective sentiment classifiers for text and image respectively. Moreover, the semantic attention is captured by LSTM and visual attention is extracted using CNN. Then these multimodal and unimodal attentions are fed to fully connected layer afterwards sentiments are predicted and then these predictions are fused for final sentiment. Almost similar performance is reported as compared with [142].

The context of discussion or talk between persons across the multimodality is the challenging task i.e. to infer the significance and association amongst the modalities. The model proposed in [122] consist of three components viz. state GRU (*s*GRU) is adapted for interlocutor state, context GRU (*c*GRU) is employed for context of discussion or talk, and emotion GRU (*e*GRU) for taking care of emotions representation. Further, significant modalities and related fused modalities based on pairwise attention for the detection of sentiment over time is utilized. In [143], a hierarchical fusion model to deal with multimodal correlations amongst texts, images, and respective social links for better analysis of sentiment is proposed. LSTMs are used in three level i.e. word-, sentence-, and paragraph-level to extract the features from text. CNN is utilized for feature extraction and weighted network are applied to find the weights of social links. The fusion of node embedding and extracted features is performed by MLP to capture sentiment. The investigation reported by this method outperforms other baseline methods.

Mattia Atzeni at el. [5] presented ensemble method based on Bi-LSTM and neural attention approach for the classification of sentiment polarity. The higher weights are considered to hidden states for important words while applying attention mechanism. Hence, word embeddings and attention mechanism are important in sentiment classification by producing original embeddings for words. In this model, Bi-LSTM is employed to fine-tune word embeddings based on sentiment polarity and MLP is adapted for attention mechanism. Moreover, proposed model is applied on humanoid robot for the prediction of sentiment related to interacting or talking person. These results are presented in European Semantic Web Conference (ESWC) 2018 Challenge, shows that the ensemble model gives better results compared to other systems in the challenge. In [18], modalities features are extracted and prominent features directed to common AffectiveSpace i.e. cluster of diverse emotions. Here, affective commonsense in English text is represented by vector space called as *AffectiveSpace*. Once, degree of emotion is predicted using fuzzy classifier and concept classification determination by RNN, then combined result is utilized for final emotion prediction. The 24 emotions can be easily predicted with significant improvement in accuracy varies from 10% to 20%. Nevertheless, sentiment prediction in conversations is another challenging task in which emotions of speaking person changes dynamically. A framework using quantum computing (QC) and LSTM with quantum theory mathematics is developed by [161]. This multimodal fusion for decision making captures the relations between each utterance. The system outperforms other models evaluated on IEMOCAP and Multimodal EmotionLines Dataset (MELD) datasets.

We show the summarization of various models in the literature for multimodal SA based on proposed model, compared models, key contributions, datasets, and results obtained in Table 6. Moreover, some suggestions and future directions to apply RNNs in sentiment analysis for multimodal data is also presented in Section 7.

7 Summary and future directions

In this article, we reviewed the latest findings of more than 150 articles on the SA and its related tasks. Depending on the input nature, two unimodal inputs i.e. *visual* and *textual* and the combination of this as a multimodal input is sketched along with the various aspects of them. We presented the challenges, applications, issues, and recent advancements in textual, visual, and multimodal SA using sequential deep neural networks viz. RNN and its architectural variants. The architectural aspects and applicability of these models for SA and related tasks have been investigated in detail. For each unimodal input i.e. textual and visual, we explore state of the arts using RNNs. We also summarized the relevant surveys/review that covers the different dimensions of SA such as challenges, its categorization on theoretical and technical basis, and various machine learning and deep learning based models for SA. On the application point of view, we presented the various applications of SA, different applications of RNNs in various domains, and lastly the detailed applications of RNNs in textual, visual, and multimodal SA.

Textual SA processes large unlabeled text using unsupervised fashion or labeled input using supervised manner, and extract huge semantic information using these deep models. In our review on textual SA, we basically focused on the various subtasks for which RNNs were employed. These subtasks include *embeddings*, *refinement of embeddings*, *classification*, *feature extractions*, *contextual information*, etc. Visual SA extracts more abstract features via deep CNN and features are extracted from a sequence of visual frames using RNN, LSTM, and its variants. The multimodal SA processes each modality separately at feature level or decision level, and finally uses any appropriate model to join these individual modalities to generate cumulative SA score or polarity. Improved results have been obtained by employing architectural variants in the deep models in textual, visual, and multimodal SA. The architectural variants of RNNs via changes in the different gating mechanisms of the cells are also experimented in the course of improvement of SA process. It may be considered as a trade-off between the cell complexity as a function of tensor operations and the model performances. In addition, we also presented a consolidated tabular summarization of the textual and multimodal SA illustrating the model, previous models with which the proposed model are compared, important characteristics or contributions of the proposed model, underlying datasets, and results obtained, provides a clear reflection of the recent trends and approaches. However, we could not perform the same for visual SA due to the comparative lesser literature present on visual SA specially.

Recent advancements in the deep models are having the scope to further optimize these tasks. Some of these may include efficient architectural designs, ensemble architectures, auto-search of optimal hyperparameters in space, improved convergence approaches in deep neural networks, etc. Efficient representation of the input has an indefinite scope of improvements in the context of these tasks under the hood of deep models. As an architectural advancement in RNNs, deeper RNNs, multidimensional RNNs, Recurrent convolutional Neural Networks, and bidirectional architectures should be rigorously explored in the SA process. LSTM is considered as a most remarkable successor of RNNs. However, we found the lack of uses of other member of the LSTM family in the SA determination in the literature such as Grid-, Differential-, Local-Global, S-, Stacked-, Matching-, and Frequency-Time LSTM.

Moreover, we identify the future directions and applications of multimodal SA for the study of emotions in the area or sub-domain like psychological studies and investigation of people where the structure of emotion can be utilized to predict the people's emotions. Also, multimodal SA can be applied in modeling the human language in the domain of language and speech processing, multiple speakers in a video, and image captioning.

Table 6 Summarization of multimodal SA using RNNs

Ref.	Compared with	Key characteristics/contributions	Datasets & Results
BDMLA [142]	Late fusion [83], CCR [150], T-LSTM embedding [149], TFN [154], Single visual model [83], Single textual model [83], Early fusion [83]	• Bi-direction attentions mechanism • Sentiment classification using multi-level correlations amongst text and image	• Flickr [12], accuracy – 84.90% • Flickr-ML, accuracy – 87.8% • Getty Images, accuracy – 86.5% • Flickr-IML [73], accuracy – 83.1%
[95]	Poria et al. [111], Zadeh et al. [154]	• GRU for context-aware feature extraction at first layer • Hierarchical feature fusion • CNN for text extraction, openSMILE for audio extraction, and 3d-CNN for video feature extraction	• CMU-MOSI, accuracy −80.0% • IEMOCAP, accuracy - 76.5%
[42]	Poria et al. [111], Zadeh et al. [154], Zadeh et al. [155]	• Multimodal multi-utterance SA • Multimodal uni-utterance SA • Multi-utterance self-attention	• CMU-MOSI, accuracy – 82.31% • CMU-MOSEI, accuracy – 79.80%
DMAF [59]	Late fusion [83], CCR [150], T-LSTM embedding [149], TFN [154], Single visual model [83], Single textual model [83], Early fusion [83]	• Multimodal attention fusion for both text and image • Unimodal attention fusion for text and image • Late fusion of unimodal and multimodal attention for sentiment prediction	• Twitter, accuracy – 76.30% • Flickr-w, accuracy – 85.9% • Getty Images, accuracy – 86.9% • Flickr-m, accuracy – 88.0%
Multilogue-Net [122]	BC-LSTM, MMMU-BA [42], Graph-MFN	• *s*GRU used for interlocutor state • *c*GRU captures context of discussion or talk • *e*GRU for emotion representation • pairwise attention mechanism	• CMU-MOSI, accuracy - 81.19% • CMU-MOSEI, accuracy - 82.10%
HDF [143]	Late fusion [83], CCR [150], T-LSTM embedding [149], TFN [154], Single visual model [83], Single textual model [83], Early fusion [83]	• Hierarchical fusion model • 3-level LSTMs applied to combine the text and image content • Network of weighted relation	• Flickr, accuracy – 85.9% • Twitter, accuracy - 76.7% • Flickr-ML, accuracy – 88.1%
[56]	Inception model	• deep neural networks combining visual analysis and NLP • LSTM for text processing	Tumblr dataset, accuracy – 72.00%
LMF [90]	SVM Deep Fusion	• Low-rank Multimodal Fusion method • Tensor flow	CMU-MOSI, MAE – 0.912 POM, MAE – 0.796

Table 6 (continued)

Ref.	Compared with	Key characteristics/contributions	Datasets & Results
	Tensor Fusion Network Multi-View LSTM		IEMOCAP, accuracy - 85.9%
TFS [154]	C-MKL SAL-CNN SVM-MD	• Tensor Fusion Network with intra-modality and inter-modality dynamics • Tensor flow	CMU-MOSI, accuracy – 77.10%
MARN [156]	SVM RF THMM	• novel neural architecture • LSTM as modality memory	CMU-MOSI, accuracy – 77.10% ICT-MMMO, accuracy – 86.30% YouTube, accuracy – 54.20% MOUD, accuracy – 81.10% IEMOCAP, accuracy – 37.00%
RAVEN [133]	SVM BC-LSTM MARN MFN RMFN LMF	• Word Representations: fine-grained structure of nonverbal behaviors with visual and acoustic pattern • LSTM computes visual and acoustic embedding	CMU-MOSI, accuracy – 78.00% IEMOCAP, accuracy – 82.00%

References

1. Abburi H, Gangashetty SV, Shrivastava M, Mamidi R Audio and Text based Multimodal Sentiment Analysis using Features Extracted from Selective Regions and Deep Neural Networks (Doctoral dissertation, International Institute of Information Technology Hyderabad).
2. Alayba AM, Palade V, England M, Iqbal R (2018) A combined cnn and lstm model for arabic sentiment analysis. In: International cross-domain conference for machine learning and knowledge extraction. Springer, Cham, pp 179–191
3. Al-Moslmi T, Omar N, Abdullah S, Albared M (2017) Approaches to cross-domain sentiment analysis: a systematic literature review. IEEE Access 5:16173–16192
4. Araque O, Barbado R, Sánchez-Rada JF, Iglesias CA (2017) Applying recurrent neural networks to sentiment analysis of spanish tweets Proc TASS 1896.
5. Atzeni M, Recupero DR (2019) Multi-domain sentiment analysis with mimicked and polarized word embeddings for humanrobot interaction. Futur Gener Comput Syst 110:984–999
6. Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2016) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75(5):2507–2525
7. Bai X (2011) Predicting consumer sentiments from online text. Decis Support Syst 50(4):732–742
8. Balamurali AR, Joshi A, Bhattacharyya P (2011) Robust sense-based sentiment classification. In: Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, pp 132-138.
9. Beel J, Gipp B, Langer S, Breitinger C (2016) Paper recommender systems: a literature survey. Int J Digit Libr 17(4):305–338
10. Bengio Y, Courville AC, Vincent P (2012) Unsupervised feature learning and deep learning: a review and new perspectives. CoRR, abs/1206.5538, 1(2012).
11. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
12. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: proceedings of the 21st ACM international conference on multimedia, pp 223-232.
13. Britz D (2015) Recurrent neural networks tutorial, part 1 – introduction to RNNs. http://www.wildml.Com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns.
14. Budiharto W, Meiliana M (2018) Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. J Big Data 5(1):51
15. Bullas J (2014) (22) social media facts and statistics you should know in 2014. Jeffbullas.com. https://www.jeffbullas.com/20-social-media-factsand-statistics-you-should-know-in-2014/. Accessed 13 Apr 2020
16. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
17. Caschera MC, Ferri F, Grifoni P (2016) Sentiment analysis from textual to multimodal features in digital environments. In: proceedings of the 8th international conference on Management of Digital EcoSystems, pp 137-144.
18. Chaturvedi I, Satapathy R, Cavallari S, Cambria E (2019) Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recogn Lett 125:264–270
19. Che Z, Purushotham S, Cho K, Sontag D, Liu Y (2018) Recurrent neural networks for multivariate time series with missing values. Sci Rep 8(1):1–12
20. Chen PC, Pavlidis T (1979) Segmentation by texture using a co-occurrence matrix and a split-and-merge algorithm. Comput Graphics Image Process 10(2):172–182
21. Chen T, Yu FX, Chen J, Cui Y, Chen YY, Chang SF (2014) Object-based visual sentiment concept analysis and application. In: Proceedings of the 22nd ACM international conference on multimedia, pp 367-376.
22. Chen R, Zhou Y, Zhang L, Duan X (2019) Word-level sentiment analysis with reinforcement learning. In: IOP conference series: materials science and engineering, pp 490(6).
23. Cheng J, Li P, Ding Z, Zhang S, Wang H (2016) Sentiment classification of chinese microblogging texts with global RNN. In: 2016 IEEE first international conference on data science in cyberspace – DSC'16, pp 653-657.
24. Cheng J, Zhang X, Li P, Zhang S, Ding Z, Wang H (2016) Exploring sentiment parsing of microblogging texts for opinion polling on chinese public figures. Appl Intell 45(2):429–442
25. Chikersal P, Poria S, Cambria E (2015) SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In: Proceedings of the 9th international workshop on semantic evaluation- SemEval'15, pp 647-651.

26. Choi M, Tani J (2017) Predictive coding for dynamic vision: development of functional hierarchy in a multiple spatio-temporal scales RNN model. In: 2017 international joint conference on neural networks – IJCNN'17, pp 657-664.
27. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
28. Dai AM, Le QV (2015) Semi-supervised sequence learning. In: Advances in neural information processing systems, pp. 3079–3087.
29. Dang NC, Moreno-García MN, De la Prieta F (2020) Sentiment analysis based on deep learning: a comparative study. Electronics 9(3):483
30. Day MY, Lin YD (2017) Deep learning for sentiment analysis on google play consumer review. In: 2017 IEEE international conference on information reuse and integration – IRI'17, pp 382-388.
31. Deng L (2014) A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Trans Signal Inf Process 3.
32. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248-255.
33. Dermouche M, Velcin J, Khouas L, Loudcher S (2014) A joint model for topic-sentiment evolution over time. In: 2014 IEEE international conference on data mining, pp 773-778.
34. Devaraj M, Piryani R, Singh VK (2016) Lexicon ensemble and lexicon pooling for sentiment polarity detection. IETE Tech Rev 33(3):332–340
35. Devika MD, Sunitha C, Ganesh A (2016) Sentiment analysis: a comparative study on different approaches. Procedia Comput Sci 87:44–49
36. Donahue J, Anne-Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.
37. Donaldson M (2017) Plitchik's Wheel of emotions–2017 Update. https://www.designwizard.com/wp-content/uploads/2017_old/09/plutchiks-modelof-emotions. Accessed 13 Apr 2020
38. Dong L, Wei F, Tan C, Tang D, Zhou M, Xu K (2014) Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: short papers), pp 49-54.
39. Eyben F, Wullmer M, Schuller BO (2018) The Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM multimedia, pp 1459-1462.
40. Fang X, Zhan J (2015) Sentiment analysis using product review data. J Big Data 2(1):5
41. Feng S, Wang Y, Liu L, Wang D, Yu G (2019) Attention based hierarchical LSTM network for context-aware microblog sentiment classification. World Wide Web 22(1):59–81
42. Ghosal D, Akhtar MS, Chauhan D, Poria S, Ekbal A, Bhattacharyya P (2018) Contextual inter-modal attention for multi-modal sentiment analysis. In: proceedings of the 2018 conference on empirical methods in natural language processing, pp 3454-3466.
43. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12).
44. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5–6):602–610
45. Graves A, Schmidhuber J (2009) Offline handwriting recognition with multidimensional recurrent neural networks. In: Advances in neural information processing systems, pp. 545–552.
46. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Chen T (2018) Recent advances in convolutional neural networks. Pattern Recogn 77:354–377
47. Guo S, Höhn S, Xu F, Schommer C (2018) PERSEUS: a personalization framework for sentiment categorization with recurrent neural network. In: International Conference on Agents and Artificial Intelligence, Funchal, pp. 94–102.
48. Hammou BA, Lahcen AA, Mouline S (2020) Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. Inf Process Manag 57(1):102122
49. Hao W, Zhang Z, Guan H (2018) Integrating both visual and audio cues for enhanced video caption. arXiv, arXiv-1711.
50. Hassan A, Mahmood A (2018) Convolutional recurrent deep learning model for sentence classification. IEEE Access 6:13949–13957
51. Hatua A, Nguyen TT, Sung AH (2017) Information diffusion on twitter: pattern recognition and prediction of volume, sentiment, and influence. In: Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, pp. 157–167.
52. Hemmatian F, Sohrabi MK (2017) A survey on classification techniques for opinion mining and sentiment analysis. Artif Intell Rev 1–51.

53. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
54. Ho SL, Xie M, Goh TN (2002) A comparative study of neural network and box-Jenkins ARIMA modeling in time series prediction. Comput Ind Eng 42(2–4):371–375
55. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
56. Hu A, Flaxman S (2019) Multimodal sentiment analysis to explore the structure of emotions. In: proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining, pp 350-358.
57. Hu S, Kumar A, Al-Turjman F, Gupta S, Seth S (2020) Reviewer credibility and sentiment analysis based user profile Modelling for online product recommendation. IEEE Access 8:26172–26189
58. Hu F, Li L, Zhang ZL, Wang JY, Xu XF (2017) Emphasizing essential words for sentiment classification based on recurrent neural networks. J Comput Sci Technol 32(4):785–795
59. Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image–text sentiment analysis via deep multimodal attentive fusion. Knowl-Based Syst 167:26–37
60. Hubert RB, Estevez E, Maguitman A, Janowski T (2018) Examining government-citizen interactions on twitter using visual and sentiment analysis. In: Proceedings of the 19th annual international conference on digital government research: governance in the data age, pp 1-10.
61. Hussein DMEDM (2018) A survey on sentiment analysis challenges. Journal of King Saud Univ Eng Sci 30(4):330–338
62. Imdb.com Traffic, Demographics and Competitors – Alexa (2018). Alexa Internet. Accessed 1 October 2018.
63. Jabreel M, Hassan F, Moreno A (2018) Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks. In: Advances in Hybridization of Intelligent Methods, pp. 39–55.
64. Ji S, Xu W, Yang M, Yu K (2012) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
65. Jiang K, Calix R, Gupta M (2016) Construction of a personal experience tweet corpus for health surveillance. In: Proceedings of the 15th workshop on biomedical natural language processing, pp 128-135.
66. Jiang K, Feng S, Song Q, Calix RA, Gupta M, Bernard GR (2018) Identifying tweets of personal health experience through word embedding and LSTM neural network. BMC Bioinf 19(8):210
67. Johnson R, Zhang T (2015) Semi-supervised convolutional neural networks for text categorization via region embedding. In: Advances in neural information processing systems, pp. 919–927.
68. Johnson R, Zhang T (2017) Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: long papers), pp 562-570.
69. Kaladevi P, Thyagarajah K (2019) Integrated CNN-and LSTM-DNN-based sentiment analysis over big social data for opinion mining. Behav Inform Technol 1-9.
70. Kalchbrenner N, Danihelka I, Graves A (2015) Grid long short-term memory. arXiv preprint arXiv: 1507.01526.
71. Kamel NS, Sayeed S, Ellis GA (2008) Glove-based approach to online signature verification. IEEE Trans Pattern Anal Mach Intell 30(6):1109–1113
72. Kansal N, Goel L, Gupta S (2020) A literature review on cross domain sentiment analysis using machine learning. Int J Artif Intell Mach Learn 10(2):43–56
73. Katsurai M, Satoh SI (2016) Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In: 2016 IEEE international conference on acoustics, speech and signal processing - ICASSP'16, pp 2837-2841.
74. Kauffmann E, Peral J, Gil D, Ferrández A, Sellers R, Mora H (2019) Managing marketing decision-making with sentiment analysis: an evaluation of the Main product features using text data mining. Sustainability 11(15):4235. https://doi.org/10.3390/su11154235
75. Kaur R, Kautish S (2019) Multimodal sentiment analysis: a survey and comparison. Int J Serv Sci Manag Eng Technol 10:38–58
76. Khan W, Malik U, Ghazanfar MA, Azam MA, Alyoubi KH, Alfakeeh AS (2019) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Comput 1–25.
77. Kim J, Kim J, Thu HLT, Kim H (2016) Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 international conference on platform technology and service - PlatCon'16, pp 1-5.
78. Kleenankandy J, Nazeer KA (2020) An enhanced tree-LSTM architecture for sentence semantic modeling using typed dependencies. arXiv preprint arXiv:2002.07775.

79. Kotzias D, Denil M, De Freitas N, Smyth P (2015) From group to individual labels using deep features. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 597-606.
80. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence.
81. Land WH, Schaffer JD (2020) The support vector machine. In: The art and science of machine intelligence. Springer, Cham, pp 45–76
82. Lauren P, Qu G, Yang J, Watta P, Huang GB, Lendasse A (2018) Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks. Cogn Comput 10(4):625–638
83. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp. 1188–1196.
84. Lee G, Jeong J, Seo S, Kim C, Kang P (2018) Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowl-Based Syst 152:70–82
85. Li L, Zhu X, Hao Y, Wang S, Gao X, Huang Q (2019) A hierarchical CNN-RNN approach for visual emotion classification. ACM Trans Multimed Comput Commun Appl 15(3s):1–7
86. Liang M, Hu X (2015) Recurrent convolutional neural network for object recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3367–3375.
87. Liao S, Wang J, Yu R, Sato K, Cheng Z (2017) CNN for situations understanding based on sentiment analysis of twitter data. Procedia Comput Sci 111:376–381
88. Liu M, Chen L, Liu B, Wang X (2015) VRCA: a clustering algorithm for massive amount of texts. In: Twenty-Fourth International Joint Conference on Artificial Intelligence, pp. 2355–2361.
89. Liu Y, Qin Z, Li P, Wan T (2017) Stock volatility prediction using recurrent neural networks with sentiment analysis. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 192–201.
90. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency LP (2018) Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064.
91. Liu T, Yu S, Xu B, Yin H (2018) Recurrent networks with attention and convolutional networks for sentence representation and classification. Appl Intell 48(10):3797–3806
92. Lu Y, Hu X, Wang F, Kumar S, Liu H, Maciejewski R (2015) Visualizing social media sentiment in disaster scenarios. In: Proceedings of the 24th international conference on world wide web, pp 1211-1215.
93. Lu K, Wu J (2019) Sentiment analysis of film review texts based on sentiment dictionary and SVM. In: Proceedings of the 2019 3rd international conference on innovation in artificial intelligence, pp 73-77.
94. Majumder N (2017) Multimodal sentiment analysis in social media using deep learning with convolutional neural networks. CIC, Instituto Politécnico Nacional.
95. Majumder N, Poria S, Peng H, Chhaya N, Cambria E, Gelbukh A (2019) Sentiment and sarcasm classification with multitask learning. IEEE Intell Syst 34(3):38–43
96. Markoff J (2012) Scientists see promise in deep-learning programs, NY Times. http://nyti.ms/sgcVec. https://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-intelligence.html. Accessed 13 Apr 2020
97. Mathews AP, Xie L, He X (2016) Senticap: generating image descriptions with sentiments. arXiv preprint arXiv:1510.01431.
98. McGurk Z, Nowak A, Hall JC (2019) Stock returns and investor sentiment: textual analysis and social media. J Econ Financ 1–28.
99. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5(4):1093–1113
100. Mezaal MR, Pradhan B, Sameen MI, Shafri M, Zulhaidi H, Yusoff ZM (2017) Optimized neural architecture for automatic landslide detection from high-resolution airborne laser scanning data. Appl Sci 7(7):730
101. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
102. Miyato T, Dai AM, Goodfellow I (2016) Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725.
103. Newberry C (28) Twitter statistics all marketers need to know in 2018. Hootsuite Blog, Retrieved October, 14, 2018.
104. Ning Y, Muthiah S, Rangwala H, Ramakrishnan N (2016) Modeling precursors for event forecasting via nested multi-instance learning. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1095-1104.

105. Nio L, Murakami K (2018) Japanese sentiment classification using bidirectional long short-term memory recurrent neural network. In: Proceedings of the 24th annual meeting Association for Natural Language Processing, pp 1119-1122.
106. Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, Ward R (2016) Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. IEEE Trans Audio Speech Lang Process (TASLP) 24(4):694–707
107. Pathak AR, Pandey M, Rautaray S (2019) Empirical evaluation of deep learning models for sentiment analysis. J Stats Manag Syst 22(4):741–752
108. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
109. Pham H, Manzini T, Liang PP, Poczos B (2018) Seq2seq2sentiment: multimodal sequence to sequence models for sentiment analysis. arXiv preprint arXiv:1807.03915.
110. 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.
111. Poria S, Chaturvedi I, Cambria E, Hussain A (2016) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th international conference on data mining – ICDM'16, pp 439-448).
112. Qiu J, Wang B, Zhou C (2020) Forecasting stock prices with long-short term memory neural network based on attention mechanism. PLoS One 15(1):e0227222. https://doi.org/10.1371/journal.pone.0227222
113. Rather AM, Agarwal A, Sastry VN (2015) Recurrent neural network and a hybrid model for prediction of stock returns. Expert Syst Appl 42(6):3234–3241
114. Ren R, Wu DD, Liu T (2018) Forecasting stock market movement direction using sentiment analysis and support vector machine. IEEE Syst J 13(1):760–770
115. Rojas-Barahona LM (2016) Deep learning for sentiment analysis. Lang Ling Compass 10(12):701–719
116. Rong X (2014) Word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.
117. Rong W, Peng B, Ouyang Y, Li C, Xiong Z (2015) Structural information aware deep semi-supervised recurrent neural network for sentiment analysis. Front Comp Sci 9(2):171–184
118. Sachan DS, Zaheer M, Salakhutdinov R (2019) Revisiting LSTM networks for semi-supervised text classification via mixed objective function. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 33:6940–6948.
119. Sak H, Senior A, Rao K, Beaufays F (2015) Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947.
120. Salehinejad H, Sankar S, Barfett J, Colak E, Valaee S (2017) Recent advances in recurrent neural networks. arXiv preprint arXiv:1801.01078.
121. Sheikh I, Illina I, Fohr D (2017) Segmentation and classification of opinions with recurrent neural networks.
122. Shenoy A, Sardana A (2020) Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation arXiv preprint arXiv:2002.08267.
123. Sheoran A, Kanojia D, Joshi A, Bhattacharyya P (2020) Recommendation chart of domains for cross-domain sentiment analysis: findings of a 20 domain study. arXiv preprint arXiv:2004.04478.
124. Sigurdsson GA, Chen X, Gupta A (2016) Learning visual storylines with skipping recurrent neural networks. In: European Conference on Computer Vision, pp. 71–88.
125. Soleymani M, Garcia D, Jou B, Schuller B, Chang SF, Pantic M (2017) A survey of multimodal sentiment analysis. Image Vis Comput 65:3–14
126. Steyn DH, Greyling T, Rossouw S, Mwamba JM (2020) Sentiment, emotions and stock market predictability in developed and emerging markets (no. 502). GLO discussion paper
127. Summarizing different types of sequence processing tasks (n.d.), https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789536089/5/ch05lvl1sec86/summarizing-different-types-of-sequence-processing-tasks
128. Tang D, Qin B, Liu T (2015) Deep learning for sentiment analysis: successful approaches and future challenges. Wiley Interdiscip Rev: Data Min Knowl Disc 5(6):292–303
129. Tani HL (2016) Characteristics of visual categorization of long-concatenated and object-directed human actions by a multiple spatio-temporal scales recurrent neural network model. arXiv preprint arXiv:1602.01921.
130. Tarasov DS (2015) Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews. In: Proceedings of the 21st international conference on computational linguistics dialog, pp 2:77-88.

131. Tomihira T, Otsuka A, Yamashita A, Satoh T (2018) What does your tweet emotion mean? Neural emoji prediction for sentiment analysis. In: proceedings of the 20th international conference on information integration and web-based applications & services, pp 289-296.
132. Vadicamo L, Carrara F, Cimino A, Cresci S, Dell'Orletta F, Falchi F, Tesconi M (2017) Cross-media learning for image sentiment analysis in the wild. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 308–317.
133. Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency LP (2019) Words can shift: dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 33:7216–7223.
134. Wang Y, Sun A, Han J, Liu Y, Zhu X (2018) Sentiment analysis by capsules. In: Proceedings of the 2018 world wide web conference, pp 1165-1174.
135. Wang J, Zhang L, Chen Y, Yi Z (2018) A new delay connection for long short-term memory networks. Int J Neural Syst 28(06):1750061
136. Wei D, Wang B, Lin G, Liu D, Dong Z, Liu H, Liu Y (2017) Research on unstructured text data mining and fault classification based on RNN-LSTM with malfunction inspection report. Energies 10(3):406
137. Wen Y, Xu A, Liu W, Chen L (2018) A wide residual network for sentiment classification. In: proceedings of the 2nd international conference on deep learning technologies, pp 7-11.
138. Werbos PJ (1990) Backpropagation through time: what it does and how to do it. Proc IEEE 78(10):1550–1560
139. Wissner-Gross A (2016) Datasets over algorithms. Edge.com. Accessed 8 January 2016.
140. Wu D, Chi M (2017) Long short-term memory with quadratic connections in recursive neural networks for representing compositional semantics. IEEE Access 5:16077–16083
141. Wu DD, Zheng L, Olson DL (2014) A decision support approach for online stock forum sentiment analysis. IEEE Trans Syst Man Cybern Syst 44(8):1077–1087
142. Xu J, Huang F, Zhang X, Wang S, Li C, Li Z, He Y (2019) Visual-textual sentiment classification with bi-directional multi-level attention networks. Knowl-Based Syst 178:61–73
143. Xu J, Huang F, Zhang X, Wang S, Li C, Li Z, He Y (2019) Sentiment analysis of social images via hierarchical deep fusion of content and links. Appl Soft Comput 80:387–399
144. Xu J, Li H, Zhou S (2015) An overview of deep generative models. IETE Tech Rev 32(2):131–139
145. Xu N, Liu AA, Wong Y, Zhang Y, Nie W, Su Y, Kankanhalli M (2018) Dual-stream recurrent neural network for video captioning. IEEE Trans Circuits Syst Video Technol 29(8):2482–2493
146. Xu N, Mao W, Chen G (2019) Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 33:371–378.
147. Yadav A, Vishwakarma DK (2019) Sentiment analysis using deep learning architectures: a review. Artif Intell Rev 1–51.
148. Yang L, Li Y, Wang J, Sherratt RS (2020) Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE Access 8:23522–23530
149. You Q, Cao L, Jin H, Luo J (2016) Robust visual-textual sentiment analysis: when attention meets tree-structured recursive neural networks. In: proceedings of the 24th ACM international conference on multimedia, pp 1008-1017.
150. You Q, Luo J, Jin H, Yang J (2016) Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: Proceedings of the Ninth ACM international conference on Web search and data mining, pp. 13–22.
151. Yu Y, Lin H, Meng J, Zhao Z (2016) Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms 9(2):41
152. Yu LC, Wu JL, Chang PC, Chu HS (2013) Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl-Based Syst 41:89–97
153. Yue B, Fu J, Liang J (2018) Residual recurrent neural networks for learning sequential representations. Information 9(3):56
154. Zadeh A, Chen M, Poria S, Cambria E, Morency LP (2017) Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.
155. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP (2018) Memory fusion network for multi-view sequential learning. arXiv preprint arXiv:1802.00927.
156. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency LP (2018) Multi-attention recurrent network for human communication comprehension. In: Thirty-Second AAAI Conference on Artificial Intelligence. (vol. 2018, pp 5642). NIH Public Access.
157. Zadeh A, Zellers R, Pincus E, Morency LP (2016) Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell Syst 31(6):82–88
158. Zaytar MA, El Amrani C (2016) Sequence to sequence weather forecasting with long short-term memory recurrent neural networks. Int J Comput Appl 143(11):7–11

159. Zhang Y, Jiang Y, Tong Y (2016) Study of sentiment classification for Chinese microblog based on recurrent neural network. Chin J Electron 25(4):601–607
160. Zhang Y, Liu Q, Song L (2018) Sentence-state LSTM for text representation. arXiv preprint arXiv: 1805.02474.
161. Zhang Y, Song D, Li X, Zhang P, Wang P, Rong L, Yu G, Wang B (2020) A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. Inform Fusion 62:14–31
162. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. Wiley Interdiscip Rev: Data Min Knowl Disc 8(4):e1253
163. Zhang XY, Yin F, Zhang YM, Liu CL, Bengio Y (2017) Drawing and recognizing chinese characters with recurrent neural network. IEEE Trans Pattern Anal Mach Intell 40(4):849–862
164. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Advances in neural information processing systems, pp. 649–657.
165. Zhao W, Guan Z, Chen L, He X, Cai D, Wang B, Wang Q (2017) Weakly-supervised deep embedding for product review sentiment analysis. IEEE Trans Knowl Data Eng 30(1):185–197
166. Zhao C, Wang S, Li D (2020) Multi-source domain adaptation with joint learning for cross-domain sentiment classification. Knowl-Based Syst 191:105254
167. Zheng J, Guo Y, Feng C, Chen H (2018) A hierarchical neural-network-based document representation approach for text classification. Math Probl Eng
168. Zhu X, Li L, Zhang W, Rao T, Xu M, Huang Q, Xu D (2019) Dependency exploitation: a unified CNN-RNN approach for visual emotion recognition. In: proceedings of the 26th international joint conference on artificial intelligence, pp 3595-3601.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.