Artificial Intelligence Review
https://doi.org/10.1007/s10462-019-09794-5

Sentiment analysis using deep learning architectures: a review

Ashima Yadav[1] · **Dinesh Kumar Vishwakarma**[1]

Abstract
Social media is a powerful source of communication among people to share their sentiments in the form of opinions and views about any topic or article, which results in an enormous amount of unstructured information. Business organizations need to process and study these sentiments to investigate data and to gain business insights. Hence, to analyze these sentiments, various machine learning, and natural language processing-based approaches have been used in the past. However, deep learning-based methods are becoming very popular due to their high performance in recent times. This paper provides a detailed survey of popular deep learning models that are increasingly applied in sentiment analysis. We present a taxonomy of sentiment analysis and discuss the implications of popular deep learning architectures. The key contributions of various researchers are highlighted with the prime focus on deep learning approaches. The crucial sentiment analysis tasks are presented, and multiple languages are identified on which sentiment analysis is done. The survey also summarizes the popular datasets, key features of the datasets, deep learning model applied on them, accuracy obtained from them, and the comparison of various deep learning models. The primary purpose of this survey is to highlight the power of deep learning architectures for solving sentiment analysis problems.

Keywords Datasets · Deep learning · Opinion mining · Review analysis · Sentiment analysis

1 Introduction

The Internet is serving as a universal and most cost-effective source of information complemented by the growth of social media. Blogs, reviews, tweets, posts, discussions on social media are scanned for extracting the opinion of people. The attitude, views, feelings, opinions constitute an essential part in analyzing the behavior of a person, which can be referred

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

Ashima Yadav
ashimayadavdtu@gmail.com

[1] Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, New Delhi 110042, India

Published online: 02 December 2019

Table 1 Applications of sentiment analysis

Applications	Refs.
Financial market prediction	Bhardwaj et al. (2015), Napitu et al. (2017), Rao and Srivastava (2012), Xu and Kešelj (2014) and Bollen et al. (2011)
Business review analysis	Zvarevashe and Olugbara (2018), Singla et al. (2017), Hedge and Padma (2017), Haque et al. (2018), Xiong et al. (2018a) and Mataoui et al. (2018)
Politics	Haselmayer and Jenny (2017), Kušen and Strembeck (2017), Kumari and Babu (2017) and Wang et al. (2012a)
Demonetization	Singh et al. (2017), Arun et al. (2017), Roy et al. (2017) and Singh et al. (2018)
Crime prediction	Wang et al. (2012b), Rosenfeld and Fornango (2008), Gerber (2014) and Azeez and Aravindhar (2015)
Disaster assessment, response, management	Ragini et al. (2018), Beigi et al. (2016), Radianti et al. (2016) and Wu and Cui (2018)

as the *sentiments*. Sentiment analysis, also known as opinion mining, deals with inspecting these sentiments directed towards any entity. Liu (2010) used the term *object* to represent the target entity mentioned in the text. An object is constituted of *components* and some set of *attributes*. For example, consider the statement "*the screen of the laptop is damaged, and it has a terrible battery life*". The object here is a laptop having *screen* and *battery* as the components. *Display quality* is an attribute of the screen, and *battery life* is the attribute of the battery. The sentiments or opinions expressed in the text are further categorized into positive, negative, neutral, or into fine-grained classification (most positive, least positive, most negative, and least negative). Hence, in the above example, negative sentiment is conveyed for the laptop object. Moreover, sentiments can be expressed in the form of text, videos, audios, emoticons, and images.

1.1 Applications of sentiment analysis

Recently, it has been observed that the number of people actively involved in social media is rapidly increasing (Facebook Statistics 2019; Twitter Statistics 2019). People are expressing their opinions in the form of reviews, comments, posts, status on various topics. As a result, a tremendous amount of data is generated on the Internet, which can be analyzed for further research. This makes sentiment analysis a popular field having ample applications. Hence, to highlight the motivation behind the sentiment analysis and to create a more profound interest of users in this area of research, we briefly discuss the famous works in the field of sentiment analysis. Table 1 lists the widespread applications of sentiment analysis.

Stock market investment constitutes a dominant part of the economy of any country. Hence, a detailed market analysis becomes a crucial part of investing money in stock market. In Bhardwaj et al. (2015), a system is developed, which fetches the live data values of Sensex and Nifty, that serves as an essential indicators of stock market. Pre-processing and feature selection are also applied to the data followed by the sentiment analysis task to get stock market status. Rao and Srivastava (2012) analyzed the relationship between the sentiment of tweets from DJIA (Dow Jones Industrial Average), NASDAQ-100, and 13 other companies and its impact on market performance. They have used Naïve Bayesian classifier for sentiment

classification. The results show that the polarity of sentiments has a substantial impact on the stock price movement. Moreover, the sentiments of the previous week strongly impact the coming week's opening and closing stock values. Xu and Kešelj (2014) proposed a method in which SVM was used for the two-stage sentiment classification process, and the dataset consisted of Tweets from the StockTwits website. In the first stage, neutral and polarized classification was performed, followed by binary (positive and negative) classification. Then a schema was designed to analyze the labeled tweets. The experimental results showed that the overnight activity of the user on StockTwits had a positive correlation to the stock trading volume of the next day. Also, it was found that the collective sentiment of after-hours (4:00 pm–9:30 am) influenced the stock movement direction of the following day.

A significant goal of sentiment analysis is to classify and analyze the reviews related to products, hotels, online booking sites, e-commerce sites, social media, etc. Haque et al. (2018) used Amazon product reviews in three domains: '*cell phone and accessories*', '*musical*', and '*electronics*'. They have classified the sentiments via Linear SVM, Multinomial Naïve Bayes, Stochastic Gradient Descent, Random Forest, Logistic regression, and Decision Tree. The best classification results were obtained by SVM with an accuracy of 94.02% on musical domain. Singla et al. (2017) have performed sentiment analysis on Amazon mobile phone reviews and in their study, they have categorized text into positive and negative polarity, and have also included sentiments of anger, anticipation, fear, joy, sadness, disgust, surprise, and trust. The classification is done through SVM resulting in an accuracy of 84.85%. Moreover, Samsung brand received the most positive feedback from customers. These results are useful for manufacturers as they can work on the feedback to improve the quality of product.

In more recent times, demonetization was announced in India, where the Government of India demonetized ₹ 500 and ₹ 1000 banknotes. Many authors have used the enormous data that originated during this period for sentiment analysis. They predicted the pros and cons of demonetization in India and analyzed the support given to the Indian government by the people. Singh et al. (2017) proposed an approach to evaluate the effect of demonetization on people all over the world. They applied a lexicon-based approach Valence Aware Dictionary and Sentiment Reasoner (VADER) for sentiment analysis on tweets extracted from Twitter. They also performed a retweet analysis in which they predicted the tweet, which could be retweeted again by using various machine learning algorithms. Experimental results show that SVM obtained finest accuracy in the case of unigram, whereas unigram-bigram feature extraction and Classification and Regression Trees (CART) gave good results for bigram feature extraction. This wide range of applications motivates us to dig deep into this area for further research.

1.2 Review methodology

The following journal databases have been explored to conduct this survey:

- Springer Link
- Science Direct
- IEEE Xplore Digital Library
- Google Scholar
- ACM Digital Library
- Wiley Online Library

We have studied over 200 research papers from the above journals, IEEE, Springer, and ACM International Conferences, and Springer Book Chapters, and have shortlisted 130 research papers on sentiment analysis, which focuses on deep learning techniques only. The search

Table 2 Count of number of reviewed articles

S. no.	Articles in journals/conference proceedings/book chapters	Count
1	Neurocomputing	11
2	Expert Systems with Applications	9
3	ACM Transactions on Asian and Low-Resource Language Information Processing	1
4	Knowledge-Based Systems	6
5	Image and Vision Computing	3
6	Journal of Parallel and Distributed Computing	1
7	Neural Processing Letters	1
8	Information Fusion	5
9	Artificial Intelligence Review	5
10	IEEE Intelligent System	2
11	Information Processing and Management	2
12	Cognitive Computation	3
13	Neural Networks	1
14	IEEE/ACM Transactions on Audio, Speech, and Language Processing	1
15	IEEE Transactions on Knowledge and Data Engineering	2
16	Journal of Computational Science	1
17	IEEE Transactions on Affective Computing	1
18	IEEE Transactions on Multimedia	1
19	IEEE Transactions on Neural Networks and Learning Systems	1
20	ACM Transactions on Information Systems (TOIS)	1
21	Journal of King Saud University—Computer and Information Sciences	1
22	Decision Support Systems	2
23	Future Generation Computer Systems	2
24	Computer Vision, Springer	1
25	Language Resources and Evaluation, Springer	1
26	Journal of the American Society for Information Science and Technology	1
27	Theoretical Computer Science	1
28	Applied Intelligence, Springer	1
29	IEEE Access	4
30	Data mining and Knowledge Discovery, Wiley	2
31	Deep Learning in Natural Language Processing, Springer	1
32	IEEE Conference Proceedings	39
33	ACM Conference Proceedings	16
	Total	130

terms or keywords used in these databases include the following: Sentiment analysis, Opinion Mining, Sentiment analysis with deep learning, opinion mining with deep learning, Online Social Networks, Social Media, Deep Learning. The distribution of articles on sentiment analysis with deep learning is shown in Table 2.

An year-wise analysis for sentiment analysis using deep learning approaches is shown in Fig. 1. From Fig. 1, it can be seen that the research in this area is increasing immensely, and from the year 2015 onwards, the number of articles published in this area are growing.

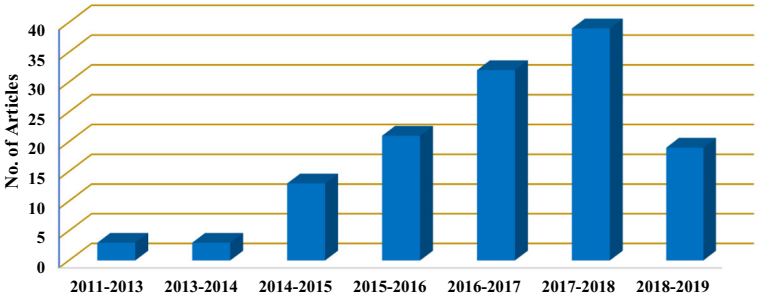


Fig. 1 Year-wise distribution of articles

Hence, our objective is to study the deep learning approaches focussed on sentiment analysis.

1.3 Earlier state-of-the-art surveys

This section discusses how our survey is different from the earlier state-of-the-art surveys. Most of the existing surveys have focused on specific areas in sentiment analysis like subjectivity detection (Chaturvedi et al. 2018) aspect extraction (Rana and Cheah 2016), social context (Sánchez-rada and Iglesias 2019), multimodal analysis (Poria et al. 2017a), information fusion (Balazs and Velásquez 2016), different languages and genre in sentiment analysis (Rani and Kumar 2019), multilingual sentiment analysis (Lo et al. 2017), and lexicon-based versus machine learning-based approaches for sentiment analysis (Hemmatian and Sohrabi 2017). Different from these surveys, this review aims to cover the significant and widespread approaches which are introduced recently in the field of sentiment analysis using deep learning. However, earlier methods are still included to give a complete view of the sentiment analysis research. The previous surveys fail to provide a detailed discussion and a comparative analysis of deep learning-based approaches for sentiment analysis, which this review aims to address. Hence, we outline the existing methods by presenting a taxonomy which explores the power of deep learning approaches and discusses how these approaches improve the performance of sentiment analysis.

This survey provides a comprehensive review of the existing literature on sentiment analysis with deep learning architecture. The significant contribution of this survey can be summarized as:

- Outlined a taxonomy for sentiment classification (Fig. 3), which includes methods using handcrafted and machine-learned feature-based approaches.
- Discuss the architecture of various deep learning-based approaches to summarize the notable work and highlight the popular deep learning-based architecture used by various researchers for sentiment analysis.
- Present the vital sentiment analysis tasks and identify why deep learning models are increasingly applied to them.
- Identify the popular sentiment analysis datasets, the accuracy obtained on them, and discuss the need for creating own corpus for sentiment analysis.

The organization of the survey is shown in Fig. 2. Section 2 discusses the taxonomy and tasks related to sentiment analysis. Section 3 gives a detailed description of the deep learning models along with an application-wise comparison of deep learning approaches, drawbacks

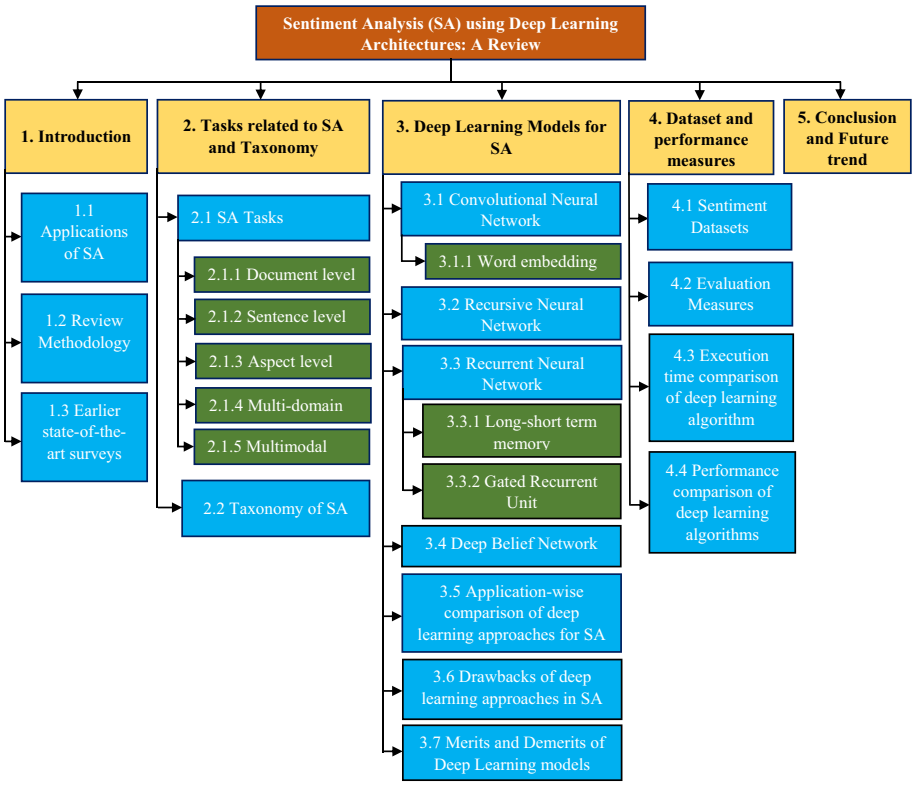


Fig. 2 Organization of the survey

of these approaches, and merits and demerits of popular deep learning models. Section 4 discusses sentiment analysis datasets and performance measures. Finally, Sect. 5 presents the conclusion and future trend.

2 Tasks related to sentiment analysis and taxonomy

This section discusses about the popular sentiment analysis tasks. Further, we have made considerable efforts to identify a taxonomy for sentiment analysis.

2.1 Sentiment analysis tasks

The sentiment analysis tasks can be categorized into two parts (Ravi and Ravi 2015): core (or major) tasks which are referred as the basic sentiment analysis tasks, and second category includes the sub-tasks that are called sub-categories of the major tasks. The core sentiment analysis tasks include document-level sentiment classification, sentence-level sentiment classification, and aspect-level sentiment classification, and sub-tasks include multi-domain sentiment classification and multimodal sentiment classification (Soleymani et al. 2017). Apart from them, various other sub-tasks of sentiment analysis, which are getting a lot of

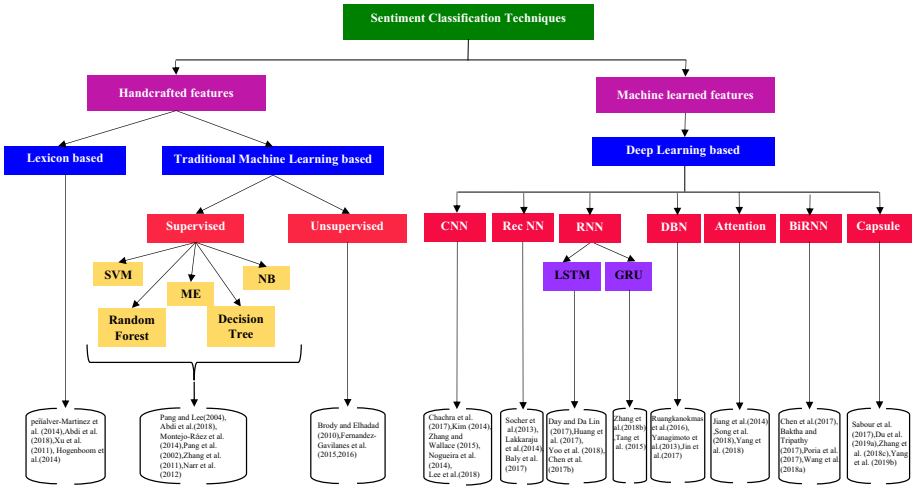


Fig. 3 Taxonomy of sentiment analysis

attention from the researchers are also identified. The various tasks and sub-tasks of sentiment analysis are discussed in the subsequent sections.

2.1.1 Document-level sentiment classification

This level of classification takes the whole document as a primary unit of information focusing on one topic or object. The document is further categorized into positive polarity or negative polarity. Thus, an overall sentiment of text can be generated. Yang et al. (2016) proposed a hierarchical attention network model that focuses on vital content for constructing the document representation. Experimental results on six popular text-based reviews demonstrate that the proposed model outperformed the state-of-the-art results by a significant margin as it can capture the insights about the structure of the document. A major challenge in document-level sentiment classification is to model long texts for generating semantic relations between sentences. This problem was handled by Huang et al. (2018), who proposed a model called SR-LSTM in which the first layer used LSTM to learn the sentence vectors, and second layer encodes the relations between the sentences. A hybrid approach of RBM and Probabilistic Neural Network (PNN) is proposed by Ghosh et al. (2017) in which RBM is used for dimensionality reduction, and PNN performs sentiment classification. The experiment is conducted in four steps: Initially, multi-domain data is collected containing reviews on Movies, Books, DVDs, Electronics, and Kitchen appliances. Next, the data is pre-processed using tokenization, stemming, and stop-word removal. The third step includes dimensionality reduction on the dataset using RBM, and finally, PNN was use for binary sentiment classification. The proposed approach gave better results on the five datasets compared to the state-of-the-art methods.

2.1.2 Sentence-level sentiment classification

The disadvantage of document-level sentiment classification is that it is difficult to extract the different polarity or sentiment about distinct entities separately. Hence, in sentence-level

sentiment classification, a sentence is classified into subjective type or objective type. A *subjective* statement expresses an opinion towards an entity. For example, "*I got a beautiful bag*", signifies positive polarity about *bag*. Hence, it is considered as a subjective statement that can be further classified into different polarities. On the other hand, factual statements are termed as *objective* statements. A statement like "*The bottle is blue in color*", displays no sentiment, so it is categorized as an objective statement.

Zhao et al. (2017) proposed a framework called Weakly-supervised Deep Embedding (WDE), which employs review ratings to train a sentiment classifier. They used CNN for constructing WDE-CNN and LSTM for constructing WDE-LSTM to extract feature vectors from review sentences. The model was evaluated on Amazon dataset from three domains: digital cameras, cell phones, and laptops. The accuracy obtained on WDE-CNN model was 87.7%, and on WDE-LSTM model was 87.9%, which shows that deep learning models gives highest accuracy as compared to baseline models. Xiong et al. (2018b) developed a model called Multi-level Sentiment-enriched Word Embedding (MSWE), which uses a Multi-layer perceptron (MLP) to model word-level sentiment information and CNN to model tweet-level sentiment information. The model also learns sentiment-specific word embeddings, and SVM is used for sentiment classification. It was evaluated on SemEval2013 dataset and Context-Sensitive Twitter (CST) dataset, which are the benchmark datasets for sentiment classification task. The F1 score obtained in the SemEval2013 dataset was 85.75, and on CST dataset was 81.34.

2.1.3 Aspect-level sentiment classification

Aspect level sentiment analysis is commonly called feature-based sentiment analysis or entity-based sentiment analysis. This sentiment analysis task includes the identification of features or aspects in a sentence (which is a user-generated review of an entity) and categorizing the features as positive or negative. The sentiment-target pairs are first identified, then they are classified into different polarities, and finally, sentiment values for every aspect are clubbed. Peng et al. (2018) studied the Chinese aspect targets at three granularity levels: radical, character, and word by proposing a model called Aspect Target Sequence Model for Single Granularity (ATSM-S). The previous work was related to processing only one aspect at a time, so they addressed this issue and presented an approach to process two aspects at a time by focusing on the aspect target itself.

Recently, attention-based LSTM mechanisms are being used for aspect-based sentiment analysis. Wang et al. (2016a) proposed an attention-based LSTM model, which can focus on different parts of a sentence when various aspects are concerned. The attention weights are computed by concatenating aspect vector into the sentence hidden representation (AE-LSTM model) or by appending aspect vector embedding into each word input vector (ATAE-LSTM model). Experimental results demonstrate that both the proposed models achieved superior performance over the baseline models, which shows that attention-based LSTM models boost the performance of aspect-based sentiment analysis models. Yu et al. (2019) proposed a framework using Bi-LSTM and multi-layer attention networks for aspect and opinion terms extraction. Al-Smadi et al. (2018) proposed a bi-LSTM with CRF model for aspect opinion target expressions (OTEs) extraction, along with aspect-based LSTM where the aspect OTEs are treated as attention expression for aspect sentiment polarity classification. Ma et al. (2018) proposed a two-step attention architecture, which attends words of the target expression along with the whole sentence. The author also applied extended LSTM, which can utilize external knowledge for developing a common-sense system for target aspect-based sentiment analysis.

The initial systems were not able to model different aspects in a sentence and do not explore the explicit position context of words. Hence, Ma et al. (2019) developed a two-stage approach that can handle the above problems. In Stage-1, position attention model is introduced for modelling the aspects and its neighboring context words. In Stage-2 multiple aspect terms within a sentence are modelled simultaneously. The most recent approach is proposed by Yang et al. (2019a), which replaces the conventional attention models with coattention mechanism by introducing a Coattention-LSTM network that can model the context-level and target-level attention alternatively by learning the non-linear representations of the target and context simultaneously. Thus, the proposed model can extract more effective sentiment features for aspect-based sentiment analysis.

2.1.4 Multi-domain sentiment classification

The word *domain* is referred as a set of documents that are related to a specific topic. Multi-domain sentiment classification focuses on transferring information from one domain to the next domain. The models are first trained in source domain; the knowledge is then transferred and explored in another domain. Dragoni and Petrucci (2017) incorporated word embeddings with a deep learning model for implementing a NeuroSent tool to build a multi-domain sentiment model. Yuan et al. (2018) proposed a Domain Attention Model (DAM) for modeling the feature-level tasks using attention mechanism for multi-domain sentiment classification. DAM is composed of two modules: domain module and sentiment module. The domain module predicts the domain in which text belongs using bi-LSTM, and sentiment module selects the important features related to the domain using another bi-LSTM with attention mechanism. The vector thus obtained from the sentiment module is fed into a softmax classifier to predict the polarity of the texts. The author used Amazon multi-domain dataset containing reviews from four domains, and Sanders Twitter Sentiment dataset containing tweets about four different IT companies. The proposed model was compared with traditional machine learning approaches, and results show that the model performed well for multi-domain sentiment classification.

2.1.5 Multimodal sentiment classification

Different people express their sentiments or opinions in different ways. Earlier, the text was considered as the primary medium to express an opinion. This is known as a unimodal approach. With the advancement of technology and science, people are now shifting towards visual (videos, images, or clips) and audio (speech) modalities to express their sentiments. Combining or fusing more than one modalities for detecting the opinion is known as multimodal sentiment analysis. Hence, researchers are now focusing on this direction for improving the sentiment classification process.

Chen et al. (2018) proposed a Weakly-Supervised Multi-modal Deep Learning (WS-MDL) model to predict multimodal sentiments for tweets. The model uses CNN and Dynamic CNN (DCNN) to calculate multimodal prediction scores and sentiment consistency scores. Due to the enormous data available on social media in different forms like videos, audios, photos for expressing sentiment on social media platforms, the conventional approach for text-based sentiment analysis was progressed into compound models of multimodal sentiment analysis. Hence, mining the opinions expressed in different modalities became a crucial approach. Poria et al. (2016a) proposed a novel methodology for merging the affective information extracted from audio, visual, and textual modalities. They discussed how different modalities

were combined together to improve the overall sentiment analysis process. Poria et al. (2018) explored three deep learning architectures for unimodal, bimodal, and multimodal (trimodal) sentiment classification. The experimental results showed that bimodal and trimodal models have shown better accuracy as compared to unimodal models, which shows the importance of using features from all the modality for enhancing the performance of sentiment analysis models. For more information on multimodal sentiment analysis, the following popular works can be referred (Soleymani et al. 2017; Chen et al. 2018; Poria et al. 2016b; Shah et al. 2016; Zhang et al. 2018a; Agarwal et al. 2019).

Table 3 discussed the deep learning models applied to various sentiment analysis (SA) tasks. From Table 3, we can conclude that deep learning models are gaining immense popularity for various SA tasks. The popular models applied to SA tasks can be summarised as:

- For document-level sentiment classification, CNN followed by LSTM has shown the highest accuracy on the various dataset.
- For sentence-level sentiment classification and aspect-level sentiment classification, researchers have majorly focused on RNN (particularly LSTM).
- For multi-domain sentiment classification, LSTM has given good results, and for multimodal sentiment classification, CNN and RNN are popular deep learning models. Hence, RNN models are the most sought-after and popular choice for sentiment analysis among researchers.
- Further, we can see that LSTM is popularly applied for text-based sentiments, and CNN models have shown good results for image sentiment. For multimodal data, CNN + LSTM followed by fusion becomes the desired approach.

However, the choice of a specific deep learning model may still depend on the various number of factors like the amount of data available, the number of hidden units (nodes) required for the problem, etc.

We found that the popular languages used for sentiment analysis are: English, Chinese, and Arabic. The detailed statistics about different types of languages used by various researchers are shown in Table 4 below.

Apart from the above-mentioned sentiment analysis tasks, we have identified various other sub-tasks that are gaining a lot of attention. This includes sarcasm detection (Halin 2017), sentiment summarization (Abdi et al. 2018), irony detection (van Hee et al. 2018), implicit sentiment detection (Chandankhede et al. 2016), temporal tagging (Hafez et al. 2017), stance detection (Krejzl et al. 2017), and emotion analysis (Hakak et al. 2017).

2.2 Taxonomy of sentiment analysis

Research in the field of sentiment analysis is taking place for several years. Initially, handcrafted features were used for various classification tasks. Some examples of handcrafted features are shown in Table 5. Lexicon based methods use handcrafted features and depend on sentiment lexicons, which are the collection of lexical units and their sentiment orientation.

On the other hand, machine-learned features can be categorized into traditional machine learning-based approaches and deep learning-based approaches. Machine learning-based methods include Support Vector Machine (SVM), Naïve Bayes (NB), Maximum Entropy (ME), Decision tree learning, and Random Forests. They are further categorized into supervised and unsupervised learning methods. A taxonomy of various approaches for sentiment analysis can be developed, as shown in Fig. 3. From the figure, it becomes evident that the number of established approaches for supervised learning is more as compared to unsu-

Table 3 Deep learning model in SA

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Uysal and Murphey (2017)	Document-level sentiment classification	English	Text	CNN, LSTM, CNN + LSTM	IMDB (Maas et al. 2011) Sentiment140 (Go et al. 2009) Nine Public Sentiment reviews (Whitehead and Yaeger 2009) Amazon Multi-domain dataset (Blitzer et al. 2007)	IMDB: 89.1 Sentiment140: 71.5 Nine Public: 77.1 Multi-domain dataset: 85.4
Huang et al. (2018)	Document-level sentiment classification	English	Text	LSTM	Yelp 2014 (https://www.yelp.com/dataset/challenge) Yelp 2015 (https://www.yelp.com/dataset/challenge) IMDB (Maas et al. 2011)	Yelp 2014: 63.9 Yelp 2015: 63.8 IMDB: 44.3

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Ghosh et al. (2017)	Document-level sentiment classification	English	Text	RBM and PNN	Movie review (MOV) (Pang and Lee 2004) Multi-domain dataset: Books (BOO), DVDs, Electronics (ELE), and Kitchen appliances (KIT) (Blitzer et al. 2007)	MOV: 80.8 BOO: 81 DVD: 81.1 ELE: 80.1 KIT: 80.2
Shi et al. (2017)	Document-level sentiment classification	Chinese	Text	Hierarchical LSTM model	Crawl weibos from Sina Weibo (http://weibo.com)	90.8
Yanagimoto et al. (2013)	Sentence classification	Japanese	Text	RBM	T&C news	–
Hassan and Mahmood (2018)	Sentence- level classification	English	Text	Joint CNN and RNN framework	IMDB (Maas et al. 2011) SSTb (Socher et al. 2013)	IMDB: 93.2 SSTb: 48.8 (fine-grained) and 89.2 (binary)
Hassan and Mahmood (2017a)	Sentence- level classification	English	Text	CNN and BRNN	IMDB (Maas et al. 2011) SSTb (Socher et al. 2013)	IMDB: 93.4 Stanford: 48.9 (Fine-grained) 89.6 (Binary)
Zhao et al. (2017)	Sentence-level classification	English	Text	CNN + LSTM	Review from Amazon on digital cameras, cell phones and laptops	WDE-CNN: 87.7 WDE-LSTM: 87.9

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Day and Da Lin (2017)	Sentence- level classification	Chinese	Text	Bidirectional-LSTM	Google Play consumer review	94
Xiong et al. (2018b)	Sentence- level classification	English	Text	MLP + CNN	SemEval2013 (Nakov et al. 2013) Context-Sensitive Twitter (CST) (Ren et al. 2016)	(F1 Score) 85.75 on SemEval 2013 And 81.34 on CST
Jin et al. (2017)	Sentence- level classification	English	Text	DBN with Delta Rule based on RBM	Stanford Twitter Sentiment dataset (STS-T) (Saif et al. 2013) STS-G (Saif et al. 2013) Health Care Reform (HCR) (Saif et al. 2013) Sentiment Strength Twitter (SST) (Saif et al. 2013) Full Twitter data (FT) (Kharde and Sonawane 2016) Game Tweets (GT)	Highest Average accuracy 67.44
Li et al. (2014)	Sentence- level classification	Chinese	Text	Recursive Neural Deep Model (RNDM)	Chinese Sentiment Treebank	90.8

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Vateekul and Koomsubha (2016)	Sentence- level classification	Thai	Text	LSTM and Dynamic Convolutional Neural Network (DCNN)	Twitter	LSTM: 75.30 DCNN: 75.35
Socher et al. (2013)	Sentence- level classification	English	Text	Recursive Neural Tensor Network (RNTN) and MV-RNN	Sentiment Treebank	RNTN: 2 way: 80.7 5 way: 87.6 MV-RNN: 2 way: 86.8 5 way: 78.7
Wu and Chi (2017)	Sentence- level classification	English	Text	qLSTM-RecNN	Stanford Sentiment Treebank (SST) (Socher et al. 2013) and SICK data set (Marelli et al. 2014)	**SST**: 86.6 (Binary) 49.4 (Fine grained) **SICK**: 87.28
Al-Smadi et al. (2017)	Aspect level sentiment classification	Arabic	Text	RNN with word embedding and SVM	Arabic Hotels' reviews dataset (Al-Smadi et al. 2016; Pontiki 2016)	RNN: 87
Peng et al. (2018)	Aspect level sentiment classification	Chinese and English	Text	LSTM + Attention mechanism	Chinese dataset (CD) SemEval2014 (SE) (Pontiki et al. 2014)	CD: 85.95 SE: 75.39
Tay et al. (2017)	Aspect level sentiment classification	English	Text	Dyadic Memory Networks	Customer reviews for laptop, restaurants, SemEval2014, Tweets from SemEval 2016, Online Debates	(F1 score) 69.2

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Lakkaraju et al. (2014)	Aspect level sentiment classification	English	Text	Joint Multi-Aspect Sentiment Model (JMAS) + RNTN	Beer reviews and camera reviews	Beer: 77.04 Camera: 81.02
Yuan et al. (2018)	Multi-domain sentiment classification	English	Text	bi-direction LSTM with attention mechanism	Amazon multi-domain dataset (Amazon) (Blitzer et al. 2007) Sanders Twitter Sentiment Dataset (Sanders)	(Average accuracy) Amazon: 87.69 Sanders: 86.32
Nozza et al. (2016)	Multi-domain sentiment classification	English	Text	Autoencoder [marginalized Stacked Denoising Autoencoder (mSDA)] and decoder	Reviews from Amazon multi-domain dataset (Blitzer et al. 2007)	87
Dragoni and Petrucci (2017)	Multi-domain sentiment classification	English	Text	RNN	Dranziera dataset (Dragoni et al. 2016)	(Average F1 score) 84.460
Poria et al. (2016b)	Multimodal sentiment classification and emotion detection	English and Spanish	Videos	Temporal CNN, RNN	MOUD dataset Youtube (Morency et al. 2011) ICT-MMMO	96.55
Chen et al. (2018)	Multimodal sentiment classification	Chinese	Image + Text + Emoticon	Visual sentiment prediction-AlexNet based CNN Textual sentiment prediction-Word2Vec and DCNN	Microblogs crawled from Sina Weibo (http://weibo.com)	69.5 (approx.)

Table 3 continued

Refs.	Sentiment analysis task	Language	Sentiment	Deep learning models	Dataset	Accuracy or F1 score (%)
Chen et al. (2017a)	Multimodal sentiment classification	English	Image + Text	Deep fusion CNN	VSO (Borth et al. 2013): Flickr Images MVSO-EN (Jou et al. 2015): the dataset includes concepts related to emotions expressed in images	VSO: 84.7 MVSO-EN: 73.7

Table 4 Count of articles according to the languages in sentiment analysis by various researchers

Language	English	Chinese	Arabic	Japanese	Vietnamese	Thai	Tibetan	Persian	Spanish	Punjabi	Chinese and English	Korean and English
Count of articles	57	18	4	2	2	1	1	1	1	1	5	2

Table 5 List of handcrafted features

Handcrafted feature	Description
PoS (Part-of-speech) (Pang and Lee 2004; Abdi et al. 2018; Poria et al. 2016c; Peñalver-Martinez et al. 2014; Montejo-Ráez et al. 2014)	PoS reads the input text and label each word with parts of speech such as nouns, adverbs, verbs, adjectives, etc.
Word n-gram, character n-gram (Pang and Lee 2004; Peñalver-Martinez et al. 2014; Pang et al. 2002; Zhang et al. 2011; Narr et al. 2012)	Word n-grams is a combination of a contiguous sequence of *n* words in the input text sequence. They are usually categorized into unigram (size 1), bigram (size 2), trigram (size 3). Similarly, character n-grams are the combination of a continuous sequence of *n* characters in the input text sequence
Hashtags (Xiong et al. 2018; Montejo-Ráez et al. 2014; Stojanovski et al. 2018)	They are majorly used for Twitter sentiment analysis by counting the positive and negative hashtag tokens in the tweet
Emoticons (Xiong et al. 2018; Montejo-Ráez et al. 2014; Narr et al. 2012; Stojanovski et al. 2018)	They are majorly used for Twitter sentiment analysis by counting the positive and negative emoticons in the tweet
Punctuation (Abdi et al. 2018; Abdi et al. 2019; Xu et al. 2011; Abbasi et al. 2008)	Occurrence of punctuation marks e.g., question mark (?), exclamation mark (!), colon (::) are considered as features
Negations (Abdi et al. 2018; Abdi et al. 2019; Zhang et al. 2017; Liu et al. 2017)	Negation words like don't, didn't, not, etc. can change the polarity of the sentence from positive to negative and vice versa
Sentiment lexicon (Abdi et al. 2018; Peñalver-Martinez et al. 2014; Abdi et al. 2019; Zhang et al. 2017)	Sentiment lexicon contains a list of words which expresses a positive or negative sentiment. The final sentiment is calculated by averaging the #positive tokens and #negative tokens. e.g., MPQA lexicon (Kiritchenko et al. 2014), SentiwordNet 3.0 (Baccianella et al. 2010), Sentiment140 lexicon (Kiritchenko et al. 2014), NRC Hashtag Sentiment Lexicon (Mohammad et al. 2013) etc.

pervised learning-based approaches. Figure 3 also gives an idea about a large number of articles published using deep learning-based approaches. These approaches include Convolution Neural Network, Recursive Neural Network, Recurrent Neural Network (which includes LSTM and GRU), Deep Belief Networks, Attention-based networks, Bi-directional Recurrent Neural Network, and capsule network.

Cambria (2016) presents a categorization for the tasks related to affective computing and sentiment analysis. The primary tasks include emotion recognition and polarity detection. The general categorization of these tasks based on existing approaches is: knowledge-based techniques, statistical techniques, and hybrid approaches. The knowledge-based techniques include several lexicons from which certain affective words are extracted for text classification. This approach is not able to handle different nuances and results in poor representation of emotions or sentiments due to the presence of certain linguistic rules. This motivated researchers to develop statistical techniques which are based on machine learning and deep learning approaches that can learn the complex features from the data, thus improving the sentiment analysis process. However, these approaches require lots of data to sufficiently train themselves. Hence, hybrid strategies that combine knowledge-based techniques and statistical techniques are popularly used for performing emotion recognition and polarity detection.

The popularity of sentiment analysis has attracted many researchers to work in this area. A lot of existing research in this field focuses on machine learning based methods. Deep learning is one of the fastest-growing areas which falls under the category of machine learning. Hence, deep learning is considered as a subject for study as it has potential benefits over other methods due to the following reasons:

- Traditional approaches, like lexicon-based approaches, use handcrafted features, which is a time-consuming and tedious process. Moreover, they are not able to generalize well for other domains or areas. Even in traditional machine learning approaches, feature engineering and feature extraction are the most time-consuming process. Hence, deep learning reduces the burden of feature design as when the network learns, it automatically creates the required features for the classification process.
- At present, a massive amount of data is being generated. According to Twitter, an average of around 6000 tweets are produced per second, which means about 200 billion tweets are being generated per year. Hence, with such a massive quantity of data, traditional machine learning-based approaches fail to perform. On the contrary, deep learning models outperform the machine learning approaches as they can be trained to learn more features with large datasets. This justifies the fact that deep learning models show improved performance than traditional machine learning models.
- The multiple layers of deep learning architecture can capture non-linear and intricate patterns in the data.
- Deep learning architectures can be adapted to other domains like Image Processing, Medical Image Segmentation, Internet of Things (IoT), Speech Recognition.

3 Deep learning models for sentiment analysis

Recent years have shown a trend of deep learning models applied in the field of natural language processing (NLP). Deep neural networks (DNNs) are made up of artificial neural networks having multiple hidden layers between the input layer and the output layer. This section provides a detailed discussion about some of the most popular deep learning mod-

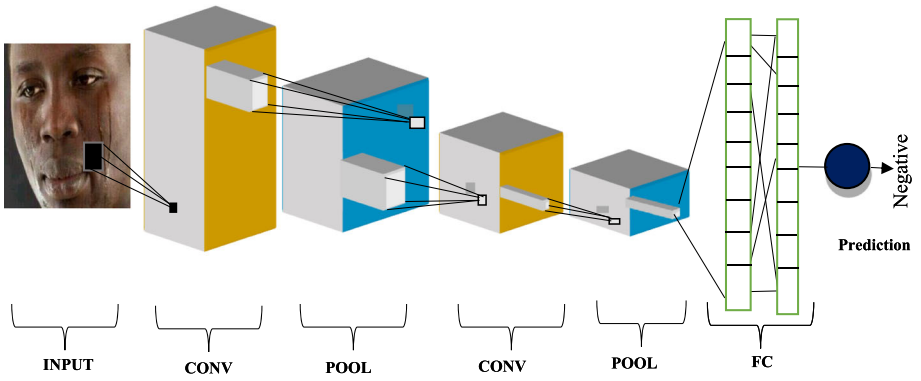


Fig. 4 CNN architecture

els like CNNs, Rec NNs, RNNs, and deep belief networks along with a brief overview of Attention-based networks, Bi-directional RNNs, and capsule networks. Further, we provide an application-wise comparison of deep learning approaches, along with their merits and demerits.

3.1 Convolutional neural networks (CNNs)

CNNs belong to the class of neural networks and have shown significant success and innovation in computer vision and image processing. The fundamental architecture of CNNis displayed in Fig. 4. As evident from the figure, CNN consists of various layers, such as the input layer, convolutional layer, pooling layer, and fully connected layer. The task of the input layer is to take the pixel value of the image as an input. Next, convolution layer (CONV) has the responsibility to produce output based on its kernel or filter values. The output obtained through a convolution operation, and Pooling Layer (POOL) is used to reduce the size of representation (dimensionality) and to speed up computation.

The most popular type of pooling is max pooling, in which maximum value from each window is taken. The Fully connected layer (FC) connects every neuron in this layer to all the activations of previous layer, as seen in ordinary neural networks. More and more researchers are actively using CNNs in the field of sentiment analysis. The most popular CNN model for sentence-level sentiment classification is the work done by Kim (2014). The author conducted an experiment with CNN built on top of pre-trained word2vec. The experimental results show that pre-trained vectors can serve as an excellent feature extractor for tasks related to NLP using deep learning. Motivated by these results, Zhang and Wallace (2015) discussed an architecture for sentence classification using one-layer CNN. They explored how the performance of a model can be affected by changing its configuration (hyperparameters, filter size, regularization parameters, etc.). Figure 5 illustrates the architecture proposed by Zhang and Wallace (2015). The tokenized sentence of length s is given as an input to the network, and it is converted into a sentence matrix by following the work of Collobert and Weston (2008) which applies a look-up table concept to generate the sentence matrix. The dimensionality of the matrix is $s * d$, where d represents dimensionality of word vectors. Hence, sentence matrix can now be treated as an input image on which convolution is performed using linear filters to generate the feature maps. The height of the filter is referred as region size of the filter. For pooling operation, 1-max pooling is performed on each feature map.

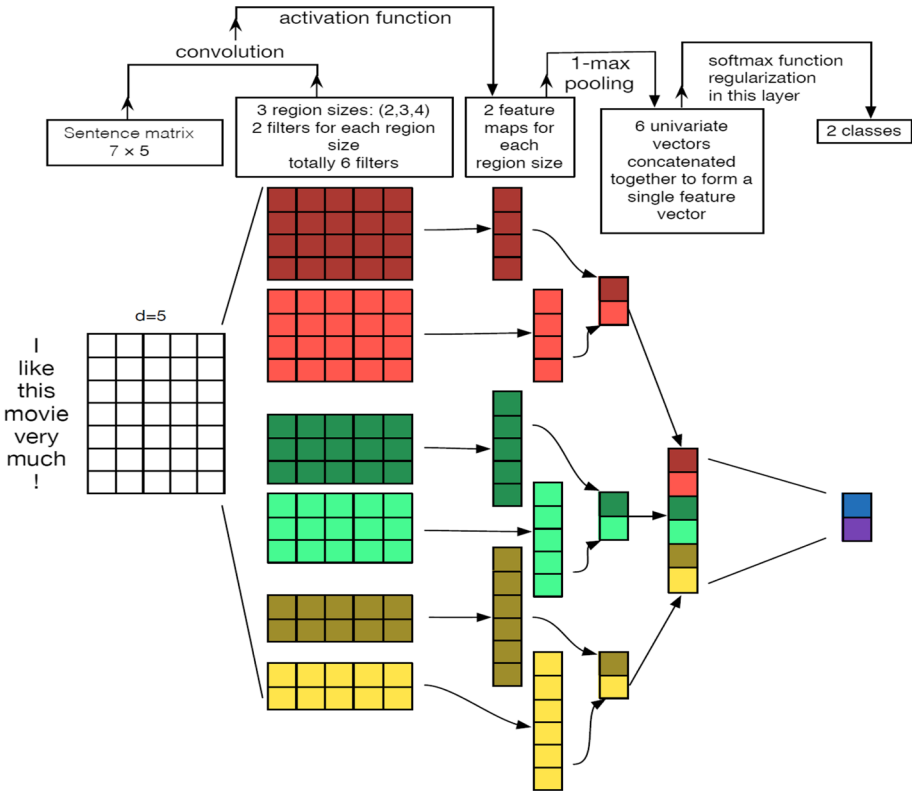


Fig. 5 CNN architecture for sentiment classification (Zhang and Wallace 2015)

3.1.1 Word embedding

Inspired by the ideas, architecture, and results of CNNs in the field of computer vision, CNNs are gaining popularity in the domain of NLP too. In NLP related tasks, an input layer consists of the matrix representation of sentences or documents, instead of the image pixels. Each row of the matrix is a vector representation of either a word or a character. These vectors are called word embedding or character embedding. The earlier approach used to represent the vocabulary of a document was one-hot encoding. The problem with this approach is that the vector size increases with the corpus size. Moreover, this encoding is not able to capture the relationship between words. Hence, word embeddings were developed as one of the most popular techniques for representing the vocabulary of a document. They contain a set of feature selection methods or a set of language models that maps the textual word into its equivalent dense and low-dimensionality vector representations. They can capture context of the word and can provide information about relation of a word with other words. Hence, meaning of a word can be predicted accurately as it can capture syntactic and semantic information about the words.

Word2Vec (Mikolov et al. 2013) is one of the famous technique for learning word embeddings as they use shallow neural network for processing a text before passing it into a deep learning algorithm. The embeddings can be obtained using Skip Gram model and Common Bag of words (CBOW) model. The CBOW model predicts the current word from surrounding

context words, whereas Skip Gram model predicts the surrounding context words from the current word. The words are mapped into a word matrix and are converted into vectors in an n-dimensional vector space by representing similar words near to each other. Similarly, Global Vectors (GloVe) (Pennington et al. 2014) generates the vector encoding of a word. The advantage of Glove model is that it can be trained quickly on more data as the implementation can be parallelized. On the other hand, instead of learning the embedding of the full word, char2vec (Cao and Rei 2016) can learn embedding associated with each character of a word.

More recently, many researchers are coming up with new approaches on word representation for sentiment analysis. The traditional word embedding methods learn word distributions that are independent of any specific task. For sentiment analysis, this can be overcome by utilizing the prior knowledge that is available in the form of sentiment labels, or opinionated words from sentiment lexicons. Li et al. (2017a) proposed a framework which combines different levels of prior knowledge into the word embeddings for sentiment analysis. Experimental results on real-world data demonstrate that the proposed word representation method improves the performance of sentiment analysis systems when compared to baseline word embedding approaches. Hao et al. (2019) proposed a novel approach by applying stochastic embeddings for cross-domain sentiment classification, which preserves similarity structures in embedding space. Yu et al. (2018) proposed a model that learns sentiment embeddings by using sentiment intensity scores from sentiment lexicons. This improves the word vectors as they are semantically and sentimentally closer to similar words.

3.2 Recursive neural networks (RecNNs)

RecNN belongs to the category of the network, that learns a directed acyclic graph structure (e.g., tree structure). The weights are shared by using them recursively over an input, which is processed in hierarchical order. The network takes a structural representation of a sentence in the form of a parse tree with word vector representations at leaves, and recursively generates parent representations in a bottom-up manner. In this way, the tokens are combined to produce representations for phrases, and finally, a complete sentence is formed. The sentence representation obtained can then be used for final sentiment classification. In this way, RecNN can learn a hierarchical structure. Since each node of the tree has a distributed feature representation associated with it, we can add a softmax layer with each parent node for computing the label probabilities. These networks are generally abbreviated as RNNs, which also stands for Recurrent Neural Network. Hence, we have abbreviated Recursive Neural Networks as Rec NN to differentiate it from Recurrent Neural Networks (RNNs).

Figure 6 shows a Rec NN in which whenever a sequence of n-gram (e.g., *This Bag has Beautiful color*), is fed into the model, each word of n-gram sequence is represented as a d-dimensional vector. Let c_i and c_j be an n-dimensional vector representation of two-child nodes, as shown in Fig. 7, having an n-dimensional parent $p_{i,j}$. The parent vector must have the same dimensions so that they can be used as an input for the next composition. The parent vector $p_{i,j}$ is fed into a softmax classifier for computing the final label probabilities.

The advantage of Rec NNs is that they are powerful enough to learn a hierarchical network. Moreover, the vector representation of the words can be immediately used as feature inputs into a softmax classifier. The disadvantage of Rec NNs is that during the training phase itself, a tree representation of the input sequence needs to be known along with the *tag* of each word (In Fig. 7, the Parts of Speech tag is labeled with every word). Moreover, the structure of every input sequence changes for each training sample, thus making them hard for training purposes.

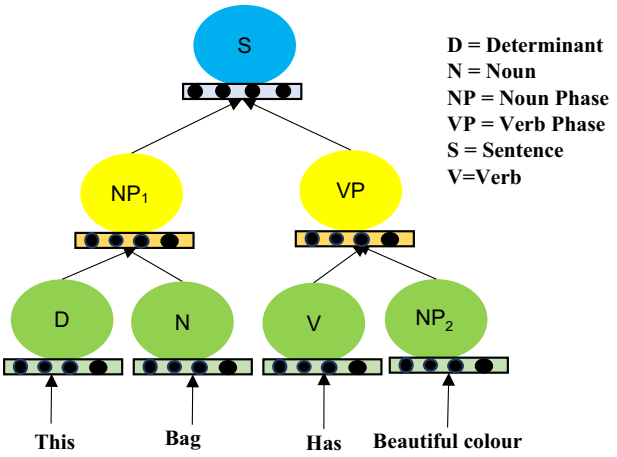
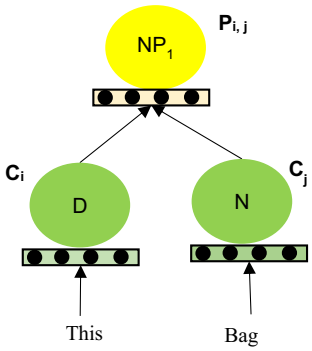


Fig. 6 Recursive neural network

Fig. 7 One Rec NN which is reproduced for each pair of input vectors



Socher et al. (2013) discussed an approach for RecNNs, as shown in Fig. 8a. A tri-gram goes as an input to the network, and it is parsed into a tree-like structure where the leaves denote a word vector. The compositional function is denoted by $g(a, p_1)$, which computes the parent nodes (vectors) by moving in a bottom-up manner. These vectors are fed into the classifiers as features. They also introduce a new corpus called Stanford Sentiment Treebank (SSTb) which will help in analysing the compositional effects of sentiments expressed in any language. They applied Recursive Neural Tensor Network (RNTN) to capture these compositional effects and increase interactions between the input vectors, as shown in Fig. 8b.

RNN, Matrix–Vector RNN (MV-RNN), and Recursive Neural Tensor Network (RNTN) are the members of the family of Recursive Neural Models. Hence, all of them follow the bottom-up approach for computing a parent vector by applying a compositionality function and considering the vector nodes as features for a classifier. The major difference between a standard RNN and RNTN model is that the latter uses a tensor-based composition function for computing vectors of higher nodes, which allows the model to have more interactions between the nodes. Hence, RNTN is more powerful as it can combine the meaning of smaller constituents of a sentence more accurately than a standard RNN. Lakkaraju et al. (2014) focused on aspect-based sentiment analysis in which various aspects of a product or service

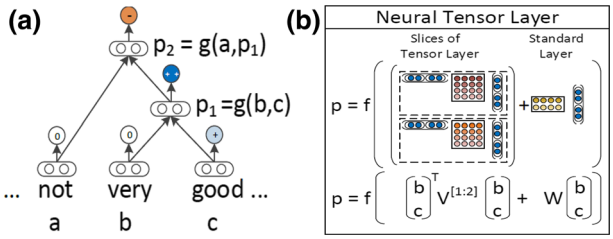


Fig. 8 a Recursive neural network model and **b** recursive neural tensor model with single layer (Socher et al. 2013)

are identified, and sentiments related to them are extracted using Joint Multi-Aspect Sentiment Model (JMAS). The results concluded that highest accuracy is shown by RNTN + JMAS followed by MV-RNN approach, which proves that simple RecNN is unable to capture the interactions between the constituents of sentences.

3.3 Recurrent neural networks (RNNs)

RNN is a variant of Rec NN and is used for modeling the sequential data. Sequential data is applied in a variety of applications. For example, in language translation, a sequence of sentence is translated from one language to another, in speech recognition an audio clip (sequence which plays over a time) is mapped into text script (sequence of words), and in video activity recognition, sequence of video frames are converted into text which describes the activity shown in video. The primary difference between RNN and Rec NN is that unlike Rec NN, RNN considers the time factor for processing the elements in a sequence. Thus, output in RNNs depends not only on the present input but also on the output computed from the previously hidden state of a network. RNNs stores the internal states of the inputs by processing each word in a sentence recurrently. Hence, to predict the next word in a sentence, RNN will store all the previous words and the relations between them.

Figure 9a shows an RNN with a loop that preserves all the information, whereas, in Fig. 9b, the unrolled version indicates that multiple copies of RNN are connected, and they communicate by passing information from one state to another. Hence, RNN with a loop signifies that it is composed of multiple copies of RNNs. The current state is computed by using the previous state value and the current input. RNNs are very popular for sentiment classification where the network classifies a piece of text into various sentiment classification tasks. The commonly used variants of RNNs are LSTM, GRU, bi-directional RNNs, and deep RNNs.

3.3.1 Long short-term memory (LSTM)

LSTM is one of the most popular variants of RNN, which possesses the capability to handle the vanishing gradient problem in standard RNN and can catch long-term dependencies. This makes them more powerful and flexible. The architecture of LSTM is shown in Fig. 10 and is comprised of forget gate f_t, input gate i_t and output gate o_t. The forget gate f_t helps in deciding which information to dump from cell at time t by taking the value of h_{t-1} (previous state information) and x_t (current input), to output a value of 0 or 1 where 0 signifies *completely dump,* and 1 signifies *completely keep.*

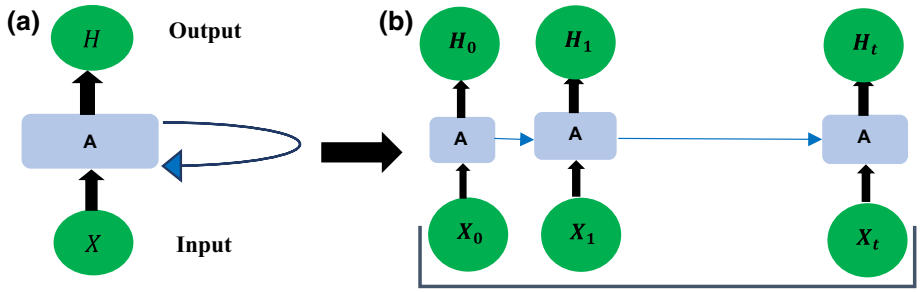


Fig. 9 **a** RNN with combined information from all the previous states and **b** unrolled version of RNN showing all the previous states

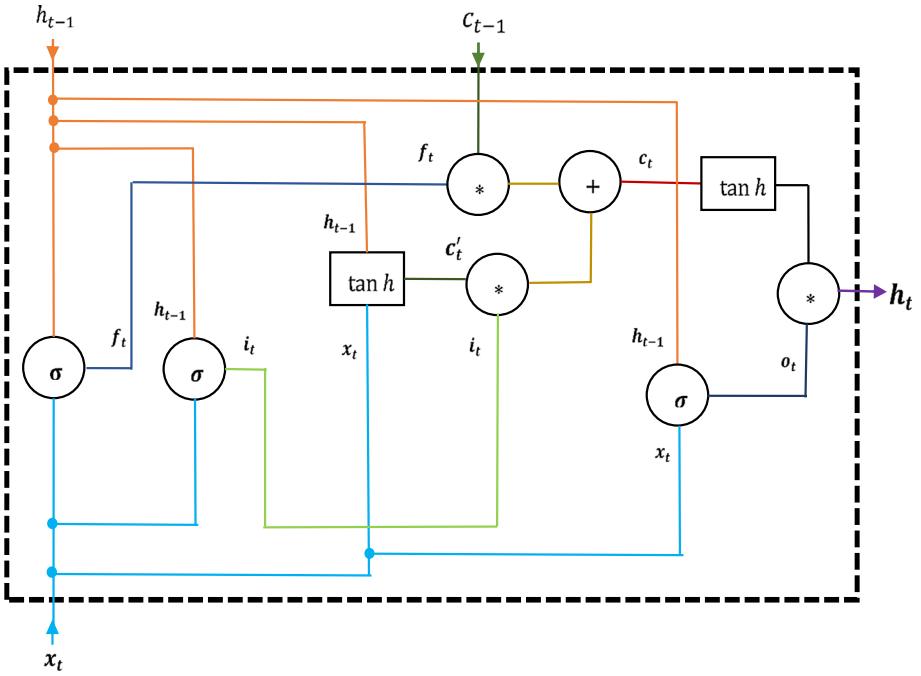


Fig. 10 Architecture of LSTM

The next step is to update the state which is done by combining these two steps: In first step, input gate i_t decides the value to be updated. In second step, *tanh* layer generates a vector with new candidate values C'_t that will be supplied to the cell's state. The old state C_{t-1} is updated to new a state C_t by multiplying it with forget gate f_t and the amount by which the state will be updated is decided by the new candidate values $i_t * C'_t$. Finally, the output gate o_t is generated by running a sigmoid layer which decides the part of the cell state that will serve as an output. Finally, for restricting the value between -1 and 1, the cell state is passed through a *tanh* layer. The resulting value is multiplied by o_t.

Researchers are also combining various deep learning techniques for sentiment classification. Huang et al. (2017) combined two popular deep learning networks called LSTM and CNN (Fig. 11) and proposed an architecture which is composed of a layer of CNN and

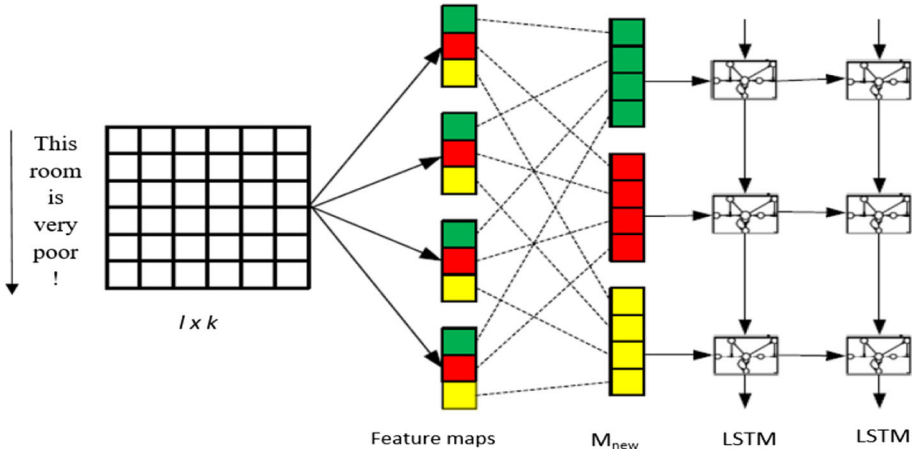


Fig. 11 Combining CNN and LSTM for sentiment classification (Huang et al. 2017)

Table 6 Comparison of work presented by Huang et al. (2017) and Hassan and Mahmood (2017b)

Refs.	Objective	Dataset	Accuracy (%)	Advantage
Huang et al. (2017)	Employed CNN to capture significant local features of the text, fed them to a two-layer LSTM model. Applied pre-trained word2vec model	Chinese Sina microblogs (binary)	87.2	The proposed model can extract context-dependent features for generating the sentence representation for sentence classification
Hassan and Mahmood (2017b)	Proposed an architecture (*ConvLstm*) which applies LSTM as a substitute of the pooling layer in CNN. Applied pre-trained word2vec model	(a) IMDB (binary) (b) SSTb (Fine-grained)	(a) 88.3 (b) 47.5	The proposed architecture can reduce the loss of local information and capture long-term dependencies

two-layer of LSTM stacked on CNN. Similarly, Hassan and Mahmood (2017b) proposed an architecture called *ConvLstm,* which again combined CNN and LSTM for classifying short texts on the top of word2vec. Table 6 shows a comparison of the work presented by both authors.

From Table 6, we observe that researchers are combining CNN + LSTM based approaches for sentence-level sentiment classification as CNN can extract the local features in the text, and LSTM can capture the long-term dependencies in sentences. Yoo et al. (2018) proposed a system called Polaris, which can analyze the sentiment trajectory of any event. Trajectory analysis can be done by classifying the contents regarding a particular event on social media according to the area where those events have occurred. The proposed model gave an F1

score of 84.1%. Chen et al. (2017b) discussed a divide-and-conquer approach for sentence-level sentiment classification. Initially, target expressions from opinionated sentences are extracted using bi-LSTM with conditional random fields (biLSTM-CRF), and sentences are classified into a non-target group, one-target group, and multi-target group. The 1d-CNNs are trained separately on each of the target groups and are used for obtaining the sentiment polarity for every type of sentence. They conducted experiments on datasets, which include movie reviews, customer product reviews, and Stanford sentiment treebank (SST-1 and SST-2) with binary and fine-grained sentiment labeled reviews along with 11 other approaches. Experimental results signifies that dividing sentences in different targets improved the performance for sentence-level sentiment analysis.

3.3.2 Gated recurrent units (GRUs)

GRUs are variants of LSTM and are considered as LSTM without an output gate. They deal with two kinds of gate: update gate and reset gate. The architecture of GRU is shown in Fig. 12. The gating mechanism of GRU is explained as follows:

(a) *Update gate* This gate defines how much information (memory) needs to be kept for the future. The input h_{t-1} denotes information from the previous $t-1$ state, x_t represents current state value, and z_t is update gate value.
(b) *Reset gate* This gate, r_t defines how much past information the network will forget.
(c) *Calculate the candidate value* The candidate value $h_t^{'}$ is calculated by taking element-wise product (denoted by, $\odot$) of set gate r_t and h_{t-1} to determine the information which needs to be removed from previous time steps, adding Ux_t where U is a parameter vector and applying *tanh* activation function on the output.
(d) *Calculate the final memory value* To get the final memory value of the current unit h_t, update gate values z_t are required to determine which information is needed from $h_t^{'}$ and which information will be required from previous value h_{t-1}.

Attention-based GRU networks are also applied to various sentiment analysis tasks like target based sentiment classification. Zhang et al. (2018b) proposed an approach to model target sentiment classification into Q&A system using Dynamic Memory Networks (DMN). The DMN consisted of four modules: Memory module consisting of input module which encodes the input sentence and uses single-layer bi-GRU to learn semantics of every word, Question module lists the questions according to different target types and Answer module shows the sentiment expressed towards the target. The memory module is comprised of multiple attention blocks and memory updating block, where multiple attention blocks consist of soft attention, attention-based GRU network, and inner attention network. The experimental results on SemEval 2014 (laptop and restaurant reviews) and Twitter dataset proved that attention-based GRU and inner attention can be used for solving the weight bias problem, thus improving target based sentiment analysis. GRUs have also been used for sentiment classification at the document-level. Tang et al. (2015a) discussed document-level sentiment classification using deep learning models. They first used CNN or LSTM to generate sentence representations from word representations and then used GRU for encoding intrinsic relations between the sentences in document representation. The experimental results showed an accuracy of 67.6% on Yelp 2015 dataset and an accuracy of 45.3% on IMDB dataset. Hence, GRU can be used to model sentence representation.

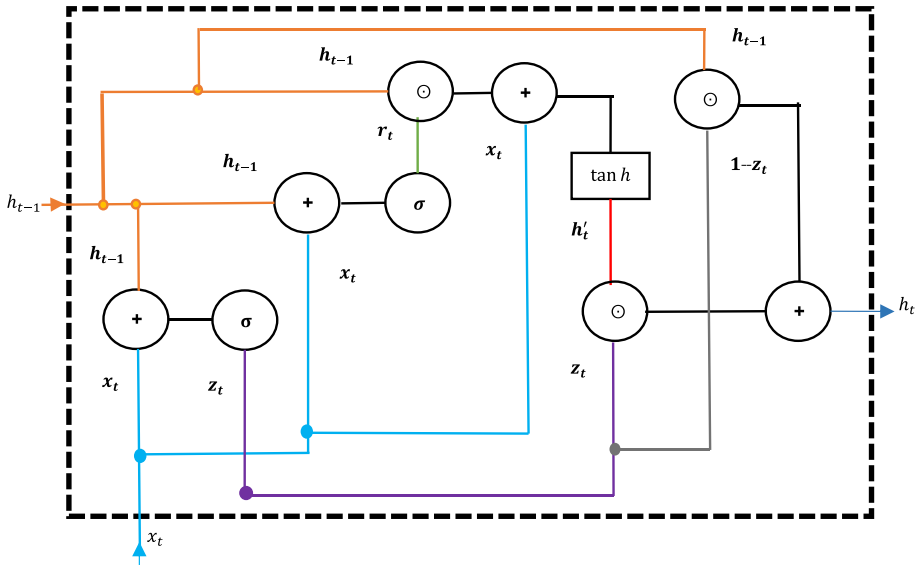


Fig. 12 Architecture of GRU

3.4 Deep belief networks (DBNs)

DBNs emanate under the type of unsupervised pre-trained networks, which also includes autoencoders and Generative Adversarial Networks. DBNs are composed of multiple layers of unsupervised networks like Restricted Boltzmann Machines (RBMs) or autoencoders as shown in Fig. 13a. Hence, various layers of RBM (Yanagimoto et al. 2013; Ruangkanokmas et al. 2016; Yuan et al. 2014) are stacked together to develop a Deep Belief Network. RBM is a stochastic neural network which consists of one layer of visible nodes $v = [V_1, V_2, V_3]$ and one layer of hidden nodes $h = [H_1, H_2]$. Each visible node is connected to hidden nodes and vice versa through an undirected connection as shown in Fig. 14a. Each hidden layer of the RBM learns the higher-level features progressively (called as pre-train phase) from the data, which are later combined for automatic feature engineering. To achieve easy learning, the restriction is made for visible nodes and hidden nodes that a visible node is not connected to other visible nodes and the same for the hidden nodes too.

A classical neural network perceptron model is shown in Fig. 14b. It is made up of three layers such as the input layer, hidden layer, and output layer. These layer consist of various nodes such as input nodes $x = [X_1, X_2, X_3]$, hidden nodes $h = [H_1, H_2, H_3, H_4]$, and output node Y and due to the use of multiple layers, it is also called a multilayer perceptron (MLP). The output of this model is computed based on input, weights, bias, and threshold parameter.

The difference between RBM and classical neural network is that an RBM has only two layers while classical neural network has three layers. RBM is an unsupervised learning (no levels) model while the classical neural network is supervised (with levels) and unsupervised both. RBM is primarily a generative model, while MLP is a discriminative model. DBNs are unsupervised networks because they are composed of unsupervised units of RBMs, while DNNs can work as a supervised or unsupervised model. DBNs are generative models, and DNNs are discriminative models. Hence, to build a DNN from a DBN, a discriminative layer

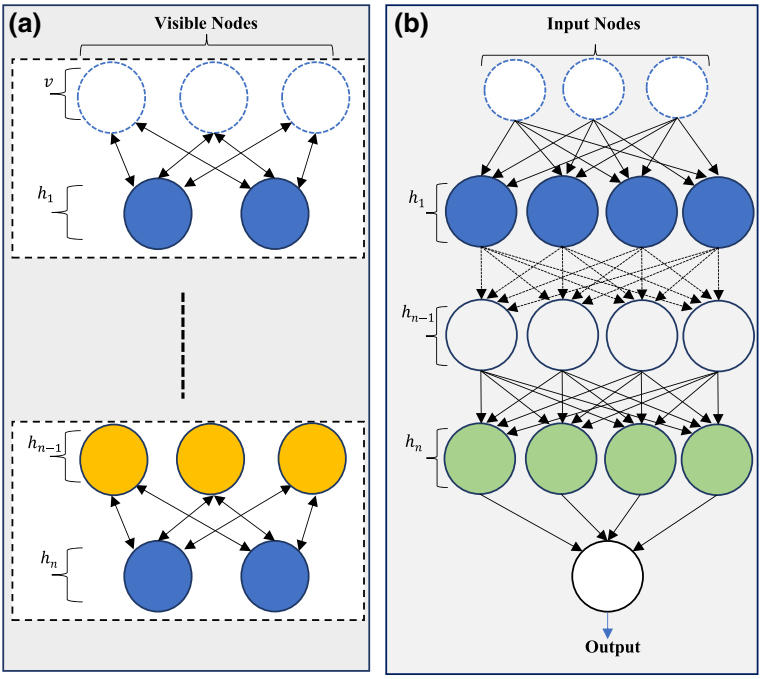


Fig. 13 **a** Deep belief network and **b** deep neural network

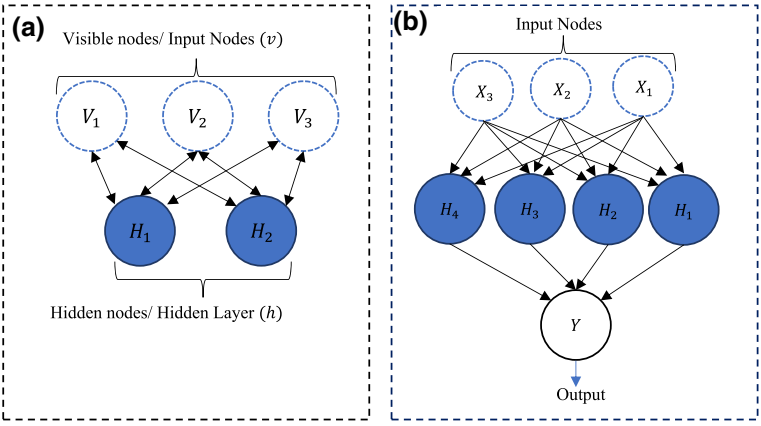


Fig. 14 **a** RBM unit and **b** classical neural network

is added on top of DBN. The key difference between DBN and DNN is in the training process as DBNs are pre-trained to reconstruct an input and then fine-tuned with backpropagation while DNN has purely supervised training with backpropagation. The pre-training of DBN is beneficial where the training set is small. Additionally, DBN supports bidirectional inter-layer communication between two different layer nodes whereas, in DNN, it is unidirectional.

Jin et al. (2017) applied DBNs with delta rule for sentiment classification on ten sentiment datasets. Delta rule uses gradient descent for fine-tuning the weights in a single layer neural

network. RBM is used to train the weights and pass them into the network through back-propagation. Experimental results show that the proposed approach (DBN with delta rule) performs better than DBN. Yanagimoto et al. (2013) developed a neural network architecture, which is composed of four layers of RBM sharing hidden layer units and visible units for estimating the similarity between the articles. The proposed approach has given good results, which shows that this architecture can be applied to the various natural language processing tasks. Ruangkanokmas et al. (2016) used DBNs with feature selection (DBNFS) for sentence classification. Initially, feature extraction is done in which pre-processing is applied to remove punctuations, stop words, and whole document is converted into feature vectors. Next, five different datasets are employed and are partitioned into an unlabeled training set, labeled training set and labeled testing set for constructing a classifier. For feature selection, Chi squared measure is used to select the most relevant features to obtain a reduced feature set. Finally, sentiment classification is done by DBNswhich outperforms the results obtained by baseline semi-supervised methods giving the best accuracy of 75.0% on electronics review dataset.

Apart from the above deep learning models, the other models which are gaining popularity in sentiment analysis are:

- *Attention-based networks* The traditional RNN approaches capture irrelevant information in the piece of information-rich text. Hence, the attention mechanism was introduced, which was inspired by the visual attention mechanism found in humans. It decides which part of the text should be focused on, rather than encoding the full sentence length. Sentiment analysis using attention-based networks is used in Yuan et al. (2018), Zhang et al. (2017, 2018b), Jiang et al. (2014), Song et al. (2018) and Yang et al. (2018).
- *Bi-directional RNN (BRNN)* A major drawback of the deep learning-based models was that they were able to learn information from previous time steps only. To overcome this issue, BRNN was introduced, which can efficiently capture the information from the future time steps for removing ambiguity and understanding the context. The module used in BRNN could be the standard RNN, LSTM, or GRU. Hence, BRNN has two types of connections: one going from left to right, in forward direction till the final time step, and other going from right to left, moving backward in time to the initial time step. Sentiment analysis using BRNN is reported in Chen et al. (2017b), Baktha and Tripathy (2017), Poria et al. (2017b) and Wang et al. (2018a).
- *Capsule networks* CNN architecture suffers from certain limitations. They were not able to model the hierarchical relationship between the local features which might misclassify objects based on their properties. The max-pooling operation in CNN results in losing certain valuable information as only the active neurons are move to the next layer. Hence, capsule networks (Sabour et al. 2017) were proposed, which can overcome these limitations. These networks consider the spatial relationships between the entities by using dynamic routing between the capsules, which is much better than max pooling operation of CNN. The dynamic routing trains the neuron vectors in capsule networks, which replaces the neuron node of the traditional neural network. The capsule networks can be trained with much less information than other neural network-based architectures. Moreover, capsule networks are getting famous in natural language processing for various text classification tasks (Du et al. 2019a; Zhao et al. 2019; Kim et al. 2019). Wang et al. (2018b) proposed RNN based capsule networks by building capsules for each sentiment category. Du et al. (2019c) proposed hybrid capsule networks for obtaining the implicit semantic information. Capsule networks are popularly used for multi-label text classification (Aly et al. 2018) and cross-domain sentiment classification (Zhang et al. 2018c; Yang et al. 2019b).

They are also applicable for aspect-level sentiment classification (Du et al. 2019b; Chen and Qian 2019). Wang et al. (2019) proposed a model in which the capsule structure can focus on each aspect category. Each capsule outputs the aspect probability and sentiment distribution on the targeted aspect. The results show that these networks can detect the words which prominently reflect the aspect category.

3.5 Application-wise comparison of deep learning approaches for sentiment analysis

This section discusses the applications of different deep learning architectures for sentiment analysis in Table 7.

3.6 Drawbacks of deep learning approaches in sentiment analysis

Although, deep learning algorithms have shown excellent outcomes and significant evolution in sentiment analysis, yet there exist some drawbacks of applying these algorithms which are stated as follows:

- Most deep learning techniques require a lot of labeled data for training to make sure that machine delivers the desired results. Hence, for sentiment analysis, we require a large corpus to properly train the deep learning model for correct prediction of the class labels. Gathering and labeling large amounts of data can be very difficult and tedious.
- Unlike traditional machine learning or lexical methods where we know what features are selected for predicting a particular sentiment, it is hard to figure out what is the actual reason for the neural network to predict a specific sentiment in a body of text just by looking at weights in different layers. This makes it difficult to get an intuition about the prediction process of the neural networks, and they behave like a "black box" to many researchers.
- Deep learning techniques like CNN requires tuning of initial parameters as a starting point. This can be seen in Stojanovski et al. (2015). Thus, the performance of the network depends upon the values of the hyperparameters of the network. Hence, deciding the optimal hyperparameter values is a challenging task.
- Due to a large number of parameters present in the deep learning models, the time taken to train them is often very large (Dufourq and Bassett 2017). Moreover, they need high performing hardware like GPUs and large RAM for better efficiency.

3.7 Merits and demerits of deep learning models

This section draws a fair comparison between different models by discussing the merits and demerits of various deep learning models in Table 8.

4 Datasets and performance measures

This section discusses the sentiment analysis dataset, evaluation measures used in sentiment analysis, execution time and performance comparison of deep learning methods.

Table 7 Application-wise comparison of deep learning approaches

Model	Refs.	Applications
CNN	Wang et al. (2016b), Tang et al. (2015b) and Do et al. (2019)	Learn the generalizable features across speakers for multimodal data (Wang et al. 2016) Capture the local semantics of n-grams of various granularities, which are proven powerful for sentiment classification (Tang et al. 2015) Identify fixed length phrases in consumer review domain (Do et al. 2019) Capture the most important n-grams feature of a sentence (Do et al. 2019)
RNN	Huang et al. (2018), Do et al. (2019) and Giachanou and Crestani (2016)	Shown great performance in Twitter domain which contains informal context and emoticons (Giachanou and Crestani 2016) Improved the performance for aspect-based sentiment analysis (Do et al. 2019) Not able to process long text due to vanishing gradient problem (Huang et al. 2018)
LSTM	Huang et al. (2018), Abdi et al. (2019) and Kraus and Feuerriegel (2019)	For text-based reviews, they are able to represent the semantic structure of a document (Kraus and Feuerriegel 2019) and model the semantic relationship in words (Abdi et al. 2019). Retain the information and order about each word in the text (Abdi et al. 2019) Able to model long text in document-level sentiment classification (Huang et al. 2018)
GRU	Zhang et al. (2018b)	BiGRU capture hidden semantics of every word in a sentence (Zhang et al. 2018c) Less computationally expensive than LSTM (Zhang et al. 2018c) Efficiently model the interaction between targeted words (Zhang et al. 2018c)
Attention	Chen (2017) and Cheng et al. (2017c)	Temporal attention + LSTM improves the sentiment prediction despite the presence of noisy acoustic and visual modalities (Chen 2017) Hierarchical attention can model the aspect information of the text, which further helps in capturing the sentiment information (Cheng et al. 2017c)
Rec NN	Do et al. (2019) and Tai et al. (2015)	Can represent multiple aspects within the text (Do et al. 2019) Learn new words and capture tree-structure of sentences (Tai et al. 2015)
DBN	Sun et al. (2016)	RBM units inside the DBN can learn the hidden structures in the input data for high-level feature representation (Sun et al. 2016)

Table 8 Merits and demerits of different deep learning models

Model	Merits	Demerits
CNN (Nogueira et al. 2014; Lee et al. 2018)	Faster and computationally less expensive as compared to RNN, LSTM, and GRU networks Fewer parameters have to be trained at the time of back propagation which makes the training significantly faster CNN combined with other networks such as LSTM has shown high accuracy Has the ability to capture relevant features from any part of the word For text, CNN models can identify relatively long sentiment/emotion phrases	Fails to preserve long-term dependency CNN ignores the long-distance dependency features that reflect syntactic and semantic information which is crucial for understanding sentences in sentiment analysis
RecNN (Socher et al. 2013; Singhal and Bhattacharyya 2016)	Powerful for learning tree-like hierarchical structure. Hence, it can be good for NLP tasks, since grammar follows a tree-like structure Have the ability to capture negation in the input sentence which is a crucial step for sentiment classification	Training can be hard as the structure changes for every training sample Model is still in its infancy stage In the case of informal data (like Tweets) which do not follow any grammatical rules, the efficiency of the model is highly affected
RNN (Hassan and Mahmood 2018)	Pays attention to the sequence of words (sequential data) in the data, so it performs better than CNN Able to store past computations Keeps the number of parameters less by weight sharing Can capture the long-distance dependency features that reflect syntactic and semantic information which are ignored by CNN Helps to preserve ordering of words in an input sequence	Not able to process very long sequences May suffer from the vanishing gradient and exploding gradients problem which makes it unfavorable for NLP tasks like sentiment analysis

Table 8 continued

Model	Merits	Demerits
LSTM (Huang et al. 2018; Peng et al. 2018; Singhal and Bhattacharyya 2016)	Can selectively forget and remember things which makes it much better for classifying sentiments Able to prevent the vanishing and exploding gradient problem in an RNN by remembering inputs from previous steps Performs much better than CNNs because of its ability to take the sequence of text in the account for predicting the sentiment For text, they can help in encoding semantics of a sentence and their representation Can be used to extract sequential information in aspect target sequence for aspect-based sentiment analysis	Computationally expensive to apply back propagation after calculating the output each time As each weight has to be trained, it can also make it harder to find an optimal solution. This also makes this network considerably slower Output produced at each time step (for every input word) needs to be reconciled to a whole sentence to get the label for the entire sentence
GRU (Verma et al. 2018; Rana 2016)	Has a fewer number of gates, hence is faster and computationally less expensive Solves the problem of vanishing and exploding gradient that is present in RNN Less complex structure than LSTM Can be useful for capturing interdependencies that exist between sentences of review document in document-level sentiment classification Have the ability to recognize emotion from a noisy speech	Does not have a memory unit, hence exposes the full hidden content without control On larger texts, LSTMs perform better than GRU because of their ability to retain memory for a longer time
DBN (Yanagimoto et al. 2013; Jin et al. 2017)	Able to learn useful information like a huge scale dimension of vocabulary from a corpus using several hidden layers	Unable to remember the previously trained task, unlike LSTM due to lack of memory Time-consuming, computationally expensive for large applications

4.1 Sentiment datasets

In order to test the performance of sentiment analysis algorithms, dataset plays a significant role. The performance is calculated in terms of prediction accuracy or F1 score and the bold letter text in Tables 9, 10 and 11 indicates the highest accuracy and model on the respective dataset. We have identified some of the popular datasets for sentiment analysis. Table 9 describes the datasets applied for sentiment analysis by describing the main features of a dataset, the deep learning models applied to them, and accuracy (or F1 score) achieved on it.

From Table 9, we conclude that some of the popular datasets which are used by various researchers in the area of sentiment analysis with deep learning models are as follows:

- *Stanford large movie review (IMDB)* (Maas et al. 2011) is a publicly available dataset consisting of 50,000 binary labeled movie reviews partitioned evenly for negative and positive reviews. A hybrid approach of RNN + CNN has shown excellent results on this dataset.
- *Yelp dataset* (Yelp Dataset 2014) consists of restaurant review labeled on a scale of 1–5 derived from the Yelp Dataset Challenge. RNN and its variants like Bi-RNN and LSTM are popularly applied to this dataset.
- *Stanford sentiment treebank (SSTb)* (Socher et al. 2013)is also a publicly available dataset containing 11,855 movie reviews from *rottentomatoes.com* website. The SSTb dataset includes five classes for classification (fine-grained sentiment classification). A hybrid approach of RNN + CNN has shown the highest results on this dataset.
- *Amazon review dataset* (Blitzer et al. 2007) is composed of product reviews from Amazon on four different types of products: books, DVDs, electronics, and kitchen appliances. Review with ratings greater than three is classified as positive and review with ratings less than three is classified as negative. A combined approach of CNN + BiLSTM + Attention has shown the highest accuracy of 87.76%.
- *CMU-MOSI dataset* (Zadeh et al. 2016) is a popular multimodal dataset which consists of 2199 opinionated utterances from 93 videos on different topics crawled from YouTube. Attention-based LSTM with dynamic fusion is the popular approach for this dataset.
- *MOUD dataset* (P'erez-Rosas et al. 2013) is another popular multimodal dataset which contains 79 videos about product reviews in Spanish. Google translate API2 is used for conversion from Spanish to English transcripts. Bi-LSTM model has shown the highest accuracy with 71.1%.
- *Getty images dataset* (You et al. 2016) consists of 588,221 labeled data that contains both images and text. Tree LSTM with Attention has given the highest accuracy on this dataset.
- *Twitter dataset* (You et al. 2016) consists of 220,000 tweets that contain both images and text. Similarly, Tree LSTM with Attention has given the highest accuracy on this dataset.
- *Twitter image dataset* (You et al. 2015) consists of 1269 image tweets labeled by five Amazon Mechanical Turk (AMT) workers. Fine-tuned CNN (GoogleNet) has given maximum accuracy on this dataset.

Hence, sentiment analysis datasets mainly comprise: Tweets, debates, messages, comments, blogs, posts, hashtags, consumer reviews on Google Play, multi-domain review about books, electronics, and kitchen appliances, reviews about restaurant, online products, hotels, and places (or locations), star ratings about products, emoticons, and images. The highest accuracy obtained on IMDB dataset is 93.2% by using CNN + RNN, and on Amazon multi-domain dataset highest accuracy obtained is around 87% by using a hybrid approach of CNN + bi-LSTM + Attention model. Hence, the RNN and its variants like LSTM, bi-RNN, bi-LSTM are giving high accuracy rates for sentiment analysis tasks, which justifies the pop-

Table 9 Description of the dataset for SA

Name of the dataset	Main feature	Deep learning models	Accuracy or F1 score (%)
Yelp 2013 (Yelp Dataset 2014)	Yelp 2013 dataset consists of restaurant review labeled on a scale of 1–5 derived from Yelp Dataset Challenge	**LSTM + KNN** (Zhou et al. 2018)	**70** (Average F1 score)
		BiLSTM (Chen et al. 2019)	56.3
		GRU (Li et al. 2019)	68.04
		Sliced RNN (Li et al. 2019)	66.02
		CNN (Tang et al. 2015)	59.6
		Bi-RNN + Attention (Zhang and Chow 2019)	**70.6**
Yelp 2014 (Yelp Dataset 2014)	Yelp 2014 dataset consists of restaurant review labeled on a scale of 1–5 derived from Yelp Dataset Challenge	GRU (Li et al. 2019)	70.90
		Sliced RNN (Li et al. 2019)	**71.60**
		CNN (Du et al. 2019d)	61.4
		Convolution + Recurrent (Du et al. 2019)	71.5
		LSTM + GRNN (Du et al. 2019)	67.1
		LSTM + KNN (Zhou et al. 2018)	69 (Average F1 score)
IMDB (Maas et al. 2011)	IMDB consists of movie reviews labeled on a scale of 1–10	LSTM + KNN (Zhou et al. 2018)	52.2 (Average F1 score)
		CNN (Lee et al. 2018)	89.57
		Joint RNN and CNN framework (Hassan and Mahmood 2018)	**93.2**
		Bi-RNN + Attention (Zhang and Chow 2019)	56.9
		Convolution + Attention (Du et al. 2019)	50.2
SSTb (Stanford Sentiment Treebank) (Socher et al. 2013)	SSTb contains labeled reviews in 5 classes for multi-class classification	LSTM (Donnelly and Roegiest 2019)	88.52
		RNN (Socher et al. 2013)	79.0
		RNTN (Socher et al. 2013)	80.7
		Bi-GRU (Baktha and Tripathy 2017)	43.81
		Joint RNN and CNN framework (Hassan and Mahmood 2018)	**89.2**

Table 9 continued

Name of the dataset	Main feature	Deep learning models	Accuracy or F1 score (%)
Amazon (Blitzer et al. 2007)	Amazon dataset contains reviews fromBooks, DVD, Electronics, and Kitchen domain from 25 products	Hierarchial attention network (HAN) (Manshu and Bing 2019) CNN (Manshu and Bing 2019) **3 layer CNN + BiLSTM + Attention** (Manshu and Bing 2019) Attention + BiLSTM (Yuan et al. 2018)	81.07 81.98 **87.76** 87.69
CMU-MOSI (Multimodal Corpus of Sentiment Intensity dataset) (Zadeh et al. 2016)	MOSI dataset consists of 2199 opinionated utterances from 93 videos on different topics crawled from YouTube	LSTM + Temporal Attention + Word level fusion (Chen 2017) LSTM + Hierarchical Fusion (Poria et al. 2017b) **Attention based LSTM + Dynamic feature fusion** (Poria et al. 2017c) Select-Additive Learning + CNN (Wang et al. 2016) Temporal Attention model (Yu et al. 2017)	75.7 80.30 **81.3** 73 75.1
MOUD (P'erez-Rosas et al. 2013)	It contains 79 videos about product reviews in Spanish. Google translate API2 is used to for conversion from Spanish to English transcripts	LSTM + Hierarchical Fusion (Poria et al. 2017) Attention based CNN and RNN + audio feature fusion (Luo et al. 2019) **Bi-LSTM** (Li and Xu 2019)	68.11 68.74 **71.1**

Table 9 continued

Name of the dataset	Main feature	Deep learning models	Accuracy or F1 score (%)
Getty Images (You et al. 2016)	It consists of 588,221 labeled dataset that contains both images and text	CNN + Distributed paragraph vector (You et al. 2016) CNN + LSTM + Attention + late fusion (Huang et al. 2019) **Tree LSTM + Attention** (Trofimova et al. 2016)	80.0 86.9 **90.2**
Twitter (You et al. 2016)	It consists of 220,000 tweets that contains both images and text	CNN + Distributed paragraph vector (You et al. 2016) CNN + LSTM + Attention + late fusion (Huang et al. 2019) **Tree LSTM + Attention** (Trofimova et al. 2016)	80.9 76.3 **96.4**
Twitter image dataset (You et al. 2015)	It consists of 1269 image tweetslabelled by five Amazon Mechanical Turk (AMT) workers	CNN (You et al. 2015) Fine-tuned CNN +oversampling (Campos et al. 2015) Fine-tuned CNN (Alexnet) (Wang et al. 2016c) **Fine-tuned CNN (GoogleNet)** (Islam and Zhang 2016) CNN +Attention (Song et al. 2018)	78.3 83.0 83.8 **86.1** 85.1

Table 10 Comparison of execution time of different deep learning algorithms

Refs.	Sentiment analysis task	Dataset (#Training sample instances)	Deep learning architecture	Execution or training time (s)	Platform	Highest accuracy (%)
Zhao et al. (2017)	Sentence classification	Review from Amazon on digital cameras, cell phones and laptops (600)	CNN **LSTM**	1800 18,000 (Execution time)	Nvidia GTX 980Ti GPU	87.9
Li et al. (2019)	Sentence classification	Yelp 2015 (808,052)	GRU **Bi-GRU** Sliced RNN Bi- Sliced RNN	1609 3176 218 440 (Training time)	Keras, NVIDIA GTX 1080Ti GPU	73.36
Li et al. (2017b)	Sentence classification	Online debates (24352), Restaurants (2614) and laptop reviews (5485) from SemEval 2014 and 2015	CNN LSTM MemNet **AttNet**	CNN:5 LSTM:200 MemNet:150 AttNet:200 (Training time)	TITAN X GPU	(Avg F-score) Debates: 52.23 Tweets: 35.34 Review: 55.93
Tay et al. (2017)	Aspect-based sentiment analysis	Customer reviews for laptop (1813), restaurants (3102), SemEval 2014 (3587), Tweets from SemEval 2016 (2771), Online Debates (24564)	**Memory NN** LSTM Attention- LSTM	Memory NN:6 LSTM:9 Attention- LSTM: 12 (Execution time)	NVIDIA GTX 1070 GPU	(Overall F1 score) 69.2
Yuan et al. (2018)	Multi-domain sentiment classification	Amazon multi-domain dataset (Amazon) (Blitzer et al. 2007) (1400) Sanders Twitter Sentiment Dataset (Sanders)	RNN GRU LSTM **BiLSTM + attention**	RNN:71 GRU:246 LSTM:310 LSTM with peephole connection:411 (Training time)	TensorFlow, NVIDIA K80 GPU	(Avg accuracy) Amazon: 87.69 Sanders: 86.32

Table 11 Performance comparison of deep learning algorithms

Refs.	Dataset	Technique/Method	Accuracy (%)	# Instances in training set
Baktha and Tripathy (2017)	Amazon health product reviews	Vanilla RNN LSTM **GRU** Bi-Vanilla RNN Bi-LSTM Bi-GRU	57.30 78.10 **83.90** 58.00 79.20 81.10	Amazon: 8000
Xu et al. (2011)	Amazon mobile phone reviews	**Multi-class SVM** CRF without interdependencies CRF with interdependencies	**61.38** 60.04 66.17	–
Rain (2013)	Amazon books, media, and kindle product reviews	Decision Tree **Naïve Bayes**	79.84 (for books) **84 (for kindle)**	–
Shaikh and Deshpande (2016)	Amazon books, camera, music product reviews	Naïve Bayes (Books) Naïve Bayes (Camera) Naïve Bayes (Music)	80 (multiword level feature) 80 (single word level feature) 80 (single word level feature)	260
Al-Smadi et al. (2017)	Arabic Hotels' Reviews (Al-Smadi et al. 2016; Pontiki 2016)	For aspect-based sentiment analysis tasks Deep RNN **SVM**	For Sentiment Polarity Identification RNN:87 **SVM:95.4**	19,226

ularity of RNN for modeling sequential data. Moreover, the combination of CNN + RNN is used for images and video-related data for visual sentiment analysis.

Most of the researchers experimented by creating their own corpus. Yang et al. (2018) created their Chinese corpus as no dataset was publicly available for target-dependent sentiment classification. Shi et al. (2017) needed user's information for user-based features. So, they crawled the weibos from Sina Weibo and labeled the weibos as positive, negative, or neutral. Due to the unavailability of publicly accessible Chinese corpus, Li et al. (2014) built a labeled parse tree corpus called "Chinese Sentiment Treebank" which is a collection of reviews about 2270 movies from a popular movie review website and made it publicly available. Hence, new datasets are needed to train the model better with a large number of records. Earlier the size of the dataset was limited to thousands of records. But now, large datasets are being formed which contain millions of records as training examples.

4.2 Evaluation measures

Most of the state-of-the-arts for sentiment analysis uses accuracy (Hassan and Mahmood 2018; Lee et al. 2018; Zhang and Chow 2019), F1 score (Xiong et al. 2018b; Dragoni and Petrucci 2017; Tay et al. 2017; Zhou et al. 2018), precision (Dragoni and Petrucci 2017; Tay

et al. 2017; Zhou et al. 2018), and recall (Dragoni and Petrucci 2017; Tay et al. 2017; Zhou et al. 2018) as performance measuring parameters. These measures are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1 score} = \frac{2 * Precision * Recall}{Precison + Recall} \tag{4}$$

where TP is the true positive, FP is the false positive, TN is the True Negative, and FN is the False Negative.

Additionally, for proper interpretation of results, there are other parameters which have been used, such as Mean Square Error, Ranking loss, Macro-averaged Mean absolute error, and RandIndex.

$$\text{Mean Square Error (MSE)} = \frac{\sum_{i=1}^{N} (Actual_i - Predicted_i)^2}{N} \tag{5}$$

MSE can be used to measure the divergence between predicted sentiment labels and actual sentiment labels (Verma et al. 2018). The sentiment models are trained by minimizing the MSE (Jiang et al. 2014; Wang et al. 2018c).

Similarly, Ranking loss measures the average distance between the true sentiment value and the predicted sentiment value for *m sentiment* classes with *n* number of test samples (Moghaddam and Ester 2010). It is defined as in Eq. (6) below:

$$\text{Ranking loss} = \sum_{i=1}^{n} \frac{\left| y_i - \hat{y}_i \right|}{m * n} \tag{6}$$

Macro-averaged Mean absolute error (MAE^M) is used for handling the imbalanced datasets (Marcheggiani and Oscar 2014; Baccianella et al. 2009).

$$\text{MAE}^M(y, \hat{y}) = \frac{1}{m} \sum_{j=1}^{k} \frac{1}{|\boldsymbol{y}_j|} \sum_{y_i \in \boldsymbol{y}_j} \left| y_i - \hat{y}_i \right| \tag{7}$$

where y is true sentiment value, $\hat{y}$ is predicted sentiment value, m is the number of sentiment labels, $\boldsymbol{y}_j$ is the subset of review corpus whose true label is j.

RandIndex is generally used for clustering problems for determining the similarity between two data clustering (Zhao et al. 2014; Xu et al. 2017; Jaffali et al. 2014). The value ranges between 0 (data clustering do not have a common pair of points) and 1 (data clustering is exactly the same). It is defined as in Eq. (8):

$$\text{RandIndex}\,(C_i, C_m) = \frac{2(x + y)}{k * (k - 1)} \tag{8}$$

where C_i *and* C_m represents the clusters produced by model i and by manual annotation m, k denotes count of detected aspects, x represents number of pairs assigned to the same cluster, and y signifies the number of pairs assigned to different clusters.

4.3 Execution time comparison of deep learning algorithms

This section compares the work proposed by various researchers by discussing the execution time or training time taken by different algorithms, as discussed in Table 10.

The architecture giving the highest accuracy is marked in Bold. As seen in Table 10, LSTM and its variants (BiLSTM, GRU) require high training and execution time as compared to other deep networks like CNN and Memory networks. However, this can be compensated by high accuracy achieved by the former. Hence, a trade-off between the execution time (or training time) with overall accuracy achieved by the network becomes a crucial task for any deep learning architecture.

4.4 Performance comparison of deep learning algorithms

This section discusses some more real text examples about Google Play reviews, Amazon reviews, and Hotel reviews to compare the performance of different deep learning algorithms on various datasets with other popular machine learning algorithms like NB, SVM to see where deep learning performs correct classifications and where it performs poorly. Table 11 explains the different techniques or methods applied to sentiment analysis.

As we see, on Amazon Review dataset, the best performance was shown by GRU with 83.90% accuracy. Apart from showing great performance in many dataset and tasks, deep learning techniques still performs poorly in many cases which can be seen in Al-Smadi et al. (2017), as we observe for aspect-based sentiment analysis, SVM has outperformed RNN in Arabic Hotel Review dataset which justifies the fact that some of the machine learning algorithms are outperforming deep learning algorithms. This may be due to the ability of SVM to extract rich hand-crafted feature set for training the model and its efficiency in binary classification tasks. Hence, determining optimal deep learning architecture is a challenging task because the performance depends on the size of dataset, type of domain or area, choosing correct hyperparameters like no. of filters, no. of hidden units, filter dimension, etc.

5 Conclusion and future trend

In this article, we reviewed the most noteworthy work on sentiment analysis using deep learning-based architectures. Firstly, we introduced sentiments and its different types, followed by their application or importance for sentiment analysis. Then, we presented a taxonomy of sentiment analysis, which included Handcrafted features based approaches, and Machine learned features based approaches along with the year-wise analysis from 2011 to 2019 for sentiment analysis including deep learning approaches only. We discussed popular deep learning models which include CNNs, Rec NNs, RNNs, LSTM, GRU, and Deep Belief Networks along with their architecture, and the important and famous work using these architectures in sentiment analysis. We have also provided a brief overview of Attention-based networks, Bi-directional RNNs, and capsule networks, which have recently gained attention of researchers. In addition, we have identified sentiment analysis tasks and discussed the deep learning models applied to them. We concluded that LSTM had given better results compared to other deep learning models. We found that different languages on which sentiment analysis is applied are: English, Chinese, Arabic, Japanese, Vietnamese, Thai, Tibetan, Persian, Spanish, and Punjabiwhereas for bilingual sentiment classification Chinese and English along with Korean and English becomes the desired approach. Finally,

we have explored primary sentiment analysis dataset, main features of the dataset, deep learning model applied to them, and the accuracy (or F1 score) obtained from dataset. We also discussed the significance of creating new datasets by various researchers, drawbacks of applying deep learning in sentiment analysis, and the merits and demerits of numerous deep learning models. We reviewed around 200 articles, and based on the detailed scrutiny of different deep learning based approaches and their state-of-the-art performances discussed in this paper; we can say that sentiment analysis using deep learning approaches is a promising research area.

We have identified and presented some datasets corresponding to various sentiment analysis tasks. The major challenge faced by many researchers is the lack of proper training datasets for different sentiment analysis tasks. Moreover, deep learning methods require huge dataset for training the model. Some researchers faced problems due to the limited size of the sentiment dataset for text retrieval tasks as the dataset contained only sentiment tags but lacked the topic tags. Fine-grained sentiment analysis can also be improved by using a large number of training examples. A strong drift is seen in transfer learning based approaches for sentiment analysis. Transfer learning can be used for transferring information from one domain to another, for building robust datasets.

The field of sentiment analysis will be highly beneficial for many domains in future such as implicit sentiment detection, spam detection, temporal tagging, and stance detection. It can be applied in the medical domain to evaluate the mental health of the patients. It may be useful by many organizations for security purposes by screening the employees to validate their integrity. Moreover, emotion and genre of a movie can also be predicted by just watching its trailer. Nowadays, people are focusing more on multimodal data for analyzing sentiments. Thus, apart from incorporating star rating and user rating in the form of text, one can opt for an exhaustive rating of products with the help of multimodal data, where the reviews of a product can be incorporated using various modalities like voice, image or emoticon.

References

Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in web forums. ACM Trans Inf Syst 26(3):12:1–12:34

Abdi A, Shamsuddin SM, Hasan S (2018) Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Syst Appl 109:66–85

Abdi A, Mariyam S, Hasan S, Piran J (2019) Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Inf Process Manag 56(4):1245–1259

Agarwal A, Yadav A, Vishwakarma DK (2019) Multimodal sentiment analysis via RNN variants. In IEEE international conference on big data, cloud computing, data science and engineering (BCD), pp 19–23

Al-Smadi M, Al-Ayyoub M, Al-Sarhan H, Jararwell Y (2016) Using aspect-based sentiment analysis to evaluate Arabic news affect on readers. In: IEEE/ACM 8th international conference on utility and cloud computing, vol 22, no 5, pp 630–649

Al-Smadi M, Qawasmeh O, Al-Ayyoub M, Jararweh Y, Gupta B (2017) Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. J Comput Sci 27:386

Al-Smadi M, Talafha B, Al-Ayyoub M, Jararweh Y (2018) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0799-4

Aly R, Remus S, Biemann C (2018) Hierarchical multi-label classification of text with capsule networks. In: Proceedings of the 35th international conference on machine learning, Sweden

Arun K, Srinagesh A, Ramesh M (2017) Twitter sentiment analysis on demonetization tweets in India using R language. Int J Comput Eng Res Trends 4(6):252–258

Azeez J, Aravindhar DJ (2015) Hybrid approach to crime prediction using deep learning. In: International conference on advances in computing, communications and informatics (ICACCI), pp 1701–1710

Baccianella S, Esuli A, Sebastiani F (2009) Multi-facet rating of product reviews. In: European conference on information retrieval. Springer, Berlin, pp 461–472
Baccianella S, Esuli A, Sebastiani F (2010) SentiwordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the seventh conference on international language resources and evaluation (LREC'10), pp 2200–2204
Baktha K, Tripathy BK (2017) Investigation of recurrent neural networks in the field of sentiment analysis. In: Proceedings of the 2017 IEEE international conference on communication and signal processing, ICCSP 2017, pp 2047–2050
Balazs JA, Velásquez JD (2016) Opinion mining and information fusion: a survey. Inf Fusion 27:95–110
Baly R, Hajj H, Habash N, Shaban KB, El-Hajj W (2017) A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic. ACM Trans Asian Low-Resour Lang Inf Process 16(4):23
Beigi G, Maciejewski R, Liu H (2016) an overview of sentiment analysis in social media and its applications in disaster relief. Stud Comput Intell 639:313–340
Bhardwaj A, Narayan Y, Vanraj P, Dutta M (2015) Sentiment analysis for indian stock market prediction using sensex and nifty. In: Procedia computer science, vol 70, pp 85–91
Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. Annu Meet Comput Linguist 45(1):440
Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
Borth D, Ji R, Chen T, Breuel T, Chang S-F (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of 21st ACM international conference on multimedia—MM'13, pp 223–232
Brody S, Elhadad N (2010) An unsupervised aspect-sentiment model for online reviews. In: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics, pp 804–812
Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107
Campos V, Salvador A, Jou B, Giró-i-nieto X (2015) Diving deep into sentiment: understanding fine-tuned CNNs for visual sentiment prediction. In: Proceedings of the 1st international workshop on affect and sentiment in multimedia. ACM, pp 57–62
Cao K, Rei M (2016) A joint model for word embedding and word morphology. In: Proceedings of the 1st workshop on representation learning for NLP, pp 18–26
Chachra A, Mehndiratta P, Gupta M (2017) Sentiment analysis of text using deep convolution neural networks. In: Tenth international conference on contemporary computing, pp 1–6
Chandankhede C, Devle P, Waskar A, Chopdekar N, Patil S (2016) ISAR: implicit sentiment analysis of user reviews. In: International conference on computing, analytics and security trends (CAST), College of Engineering Pune, India, pp 357–361
Chaturvedi I, Cambria E, Welsch RE, Herrera F (2018) Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf Fusion 44:65–77
Chen M (2017) Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM international conference on multimodal interaction. ACM, pp 163–171
Chen Z, Qian T (2019) Transfer capsule network for aspect level sentiment classification. In: Proceedings oft he 57th annual meeting of the Association for Computational Linguistics, pp 547–556
Chen X, Wang Y, Liu Q (2017a) Visual and textual sentiment analysis using deep fusion convolutional neural networks. arXiv preprint arXiv:1711.07798
Chen T, Xu R, He Y, Wang X (2017b) Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Syst Appl 72:221–230
Cheng J, Zhao S, Zhang J, King I, Zhang X, Wang H (2017c) Aspect-level sentiment classification with HEAT (hierarchical attention) network. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 97–106
Chen F, Ji R, Su J, Cao D, Gao Y (2018) Predicting microblog sentiments via weakly supervised multimodal deep learning. IEEE Trans Multimed 20(4):997–1007
Chen B et al (2019) Embedding logic rules into recurrent neural networks. IEEE Access 7:14938–14946
Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on machine learning. ACM, pp 160–167
Day MY, Da Lin Y (2017) Deep learning for sentiment analysis on google play consumer review. In: Proceedings of 2017 IEEE international conference on information reuse and integration, IRI, pp 382–388
Do HH, Prasad PWC, Maag A, Alsadoon A (2019) Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl 118:272–299
Donnelly J, Roegiest A (2019) On interpretability and feature representations: an analysis of the sentiment neuron. In: European conference on information retrieval. Springer, Cham, pp 795–802

Dragoni M, Petrucci G (2017) A neural word embeddings approach for multi-domain sentiment analysis. IEEE Trans Affect Comput 8(4):457–470
Dragoni M, Tettamanzi AGB, Pereira CDC (2016) DRANZIERA: an evaluation protocol for multi-domain opinion mining. In: Tenth international conference on language resources and evaluation, LREC, pp 267–272
Du C et al (2019a) Investigating capsule network and semantic feature on hyperplanes for text classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, pp 456–465
Du C et al (2019b) Capsule network with interactive attention for aspect-level sentiment classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, pp 5492–5501
Du Y, Zhao X, He M, Guo W (2019c) A novel capsule based hybrid neural network for sentiment classification. IEEE Access 7:39321–39328
Du J, Gui L, He Y, Xu R, Wang X (2019d) Convolution-based neural attention with applications to sentiment classification. IEEE Access 7:2169–3536
Dufourq E, Bassett BA (2017) EDEN: evolutionary deep networks for efficient machine learning. In: IEEE pattern recognition association of South Africa and robotics and mechatronics international conference, pp 110–115
Facebook Statistics (2019). https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
Fernández-Gavilanes M, Alvarez-López T, Juncal-Martínez J, Costa-Montenegro E, González-Castã FJ (2015) GTI: an unsupervised approach for sentiment analysis in twitter. In: Proceedings of 9th international workshop on semantic evaluation (SemEval 2015), pp 533–538
Fernández-Gavilanes M, Álvarez-López T, Juncal-Martínez J, Costa-Montenegro E, Javier González-Castaño F (2016) Unsupervised method for sentiment analysis in online texts. Expert Syst Appl 58:57–75
Gerber MS (2014) Predicting crime using Twitter and kernel density estimation. Decis Support Syst 61(1):115–125
Ghosh R, Ravi K, Ravi V (2017) A novel deep learning architecture for sentiment classification. In: 3rd International conference on recent advances in information technology|RAIT-2016|, pp 3–8
Giachanou A, Crestani F (2016) Like it or not: a survey of twitter sentiment analysis methods. ACM Comput Surv 49(2):28:3–28:40
Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224 N Proj Rep Stanf 1(12):1–6
Hafez G, Ismail R, Karam O (2017) Temporal sentiment analysis and time tags for opinions. In: The 8th IEEE international conference on intelligent computing and information systems (ICICIS 2017), pp 373–378
Hakak NM, Mohd M, Kirmani M, Mohd M (2017) Emotion analysis: a survey. In: International conference on computer, communications and electronics, COMPTELIX 2017, pp 397–402
Halin AA (2017) The importance of multimodality in sarcasm detection for sentiment analysis. In: IEEE 15th student conference on research and development (SCOReD), pp 56–60
Hao Y, Mu T, Hong R, Wang M, Liu X, Goulermas JY (2019) Cross-domain sentiment encoding through stochastic word embedding. IEEE Trans Knowl Data Eng 1–15
Haque TU, Saber NN, Shah FM (2018) Sentiment analysis on large scale amazon product reviews. In: IEEE international conference on innovative research and development (ICIRD), pp 1–6
Haselmayer M, Jenny M (2017) Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Qual Quant 51(6):2623–2646
Hassan A, Mahmood A (2017a) Efficient deep learning model for text classification based on recurrent and convolutional layers. In: 16th IEEE international conference on machine learning and applications (ICMLA), pp 1108–1113
Hassan A, Mahmood A (2017b) Deep learning approach for sentiment analysis of short texts. In: 3rd International conference on control, automation and robotics (ICCAR), pp 705–710
Hassan A, Mahmood A (2018) Convolutional recurrent deep learning model for sentence classification. IEEE Access 6:2169–3536
Hedge Y, Padma SK (2017) Sentiment analysis using random forest ensemble for mobile product reviews in Kannada. In: IEEE 7th international advance computing conference
Hemmatian F, Sohrabi M (2017) A survey on classification techniques for opinion mining and sentiment analysis. Artif Intell Rev 2017:1–51
Hogenboom A, Heerschop B, Frasincar F, Kaymak U, De Jong F (2014) Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decis Support Syst 62:43–53
Huang Q, Chen R, Zheng X, Dong Z (2017) Deep sentiment representation based on CNN and LSTM. In: Proceedings of 2017 international conference on green informatics, ICGI 2017, pp 30–33

Huang W, Rao G, Feng Z, Cong Q (2018) LSTM with sentence representations for document-level sentiment classification. Neurocomputing 308:49

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

Islam J, Zhang Y (2016) Visual sentiment analysis for social images using transfer learning approach. In: IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom) (BDCloud-SocialCom-SustainCom), pp 124–130

Jaffali S, Jamoussi S, Ben Hamadou A (2014) Grouping like-minded users based on text and sentiment analysis. In: International conference on computational collective intelligence. Springer, Cham, pp 83–93

Jiang M, Wang J, Lan M, Wu Y (2014) An effective gated and attention-based neural network model for fine-grained financial target-dependent sentiment analysis. Int Conf Knowl Sci Eng Manag 214:42–54

Jin Y, Zhang H, Du D (2017) Improving deep belief networks via delta rule for sentiment classification. In: Proceedings of 2016 IEEE 28th international conference on tools with artificial intelligence, ICTAI 2016, pp 410–414

Jou B, Chen T, Pappas N, Redi M, Topkara M, Chang SF (2015) Visual affect around the world: a large-scale multilingual visual sentiment ontology. In: Proceedings of the 23rd ACM international conference on multimedia. ACM, pp 159–168

Kharde VA, Sonawane SS (2016) Sentiment analysis of twitter data: a survey of techniques. Int J Comput Appl 139(11):975–8887

Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), October 25–29, Doha, Qatar, pp 1746–1751

Kim J, Jang S, Park E, Choi S (2019) Text classification using capsules. Neurocomputing 118:247–261

Kiritchenko S, Zhu X, Mohammad S (2014) Sentiment analysis of short informal texts. J Artif Intell Res 50:723–762

Kraus M, Feuerriegel S (2019) Sentiment analysis based on rhetorical structure theory: learning deep neural networks from discourse trees. Expert Syst Appl 118:65–79

Krejzl P, Hourová B, Steinberger J (2017) Stance detection in online discussions. arXiv preprint arXiv:1701.00504

Kumari S, Babu CN (2017) Real time analysis of social media data to understand people emotions towards national parties. In: 8th International conference on computing, communication and networking technologies (ICCCNT), pp 1–6

Kušen E, Strembeck M (2017) Politics, sentiments, and misinformation: an analysis of the Twitter discussion on the 2016 Austrian presidential elections. Online Soc Netw Media 5:37–50

Lakkaraju H, Socher R, Manning CD (2014) Aspect specific sentiment analysis using hierarchical deep learning. In: NIPS workshop on deep learning and representation learning, pp 1–9

Lee G, Jeong J, Seo S, Kim CY, Kang P (2018) Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowl Based Syst 152:70–82

Li H, Xu H (2019) Video-based sentiment analysis with hvnLBP-TOP feature and bi-LSTM. In: Association for the Advancement of Artificial Intelligence (AAAI)

Li C, Xu B, Wu G, He S, Tian G, Hao H (2014) Recursive deep learning for sentiment analysis over social data. In: Proceedings of 2014 IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology–workshops, WI-IAT 2014, pp 180–185

Li Y, Pan Q, Yang T, Wang S, Tang J, Cambria E (2017a) Learning word representations for sentiment analysis. Cognit Comput 9(6):843–851

Li C, Guo X, Mei Q (2017b) Deep memory networks for attitude identification. In: Proceedings of the tenth ACM international conference on web search and data mining, WSDM 2017, Cambridge, United Kingdom, pp 671–680

Li B, Cheng Z, Xu Z, Ye W, Lukasiewicz T, Zhang S (2019) Long text analysis using sliced recurrent neural networks with breaking point information enrichment. In: Proceedings of the 2019 IEEE international conference on acoustics, speech and signal processing, ICASSP 2019, Brighton, UK, vol 124, pp 51–60

Liu B (2010) Sentiment analysis and subjectivity. In: Handbook of natural language processing, vol 1, pp 1–38

Liu Y, Bi J-W, Fan Z-P (2017) Ranking products through online reviews: a method based on sentiment analysis technique and intuitionistic fuzzy set theory. Inf Fusion 36:149–161

Lo S, Cambria E, Chiong R, Cornforth D (2017) Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artif Intell Rev 48(4):499–527

Luo Z, Xu H, Chen F (2019) Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. In: Proceedings of the AAAI-19 workshop on affective content analysis, Honolulu, USA

Ma Y, Peng H, Khan T, Cambria E, Hussain A (2018) Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cognit Comput 10:639–650

Ma X, Zeng J, Peng L, Fortino G, Zhang Y (2019) Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. Futur Gener Comput Syst 93:304–311

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of 49th annual meeting of the Association for Computational Linguistics: Human Language and Technology, pp 142–150

Manshu Y, Bing W (2019) Adding prior knowledge in hierarchical attention neural network for cross domain sentiment classification. IEEE Access 7:2169–3536

Marcheggiani D, Oscar T (2014) Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In: European conference on information retrieval. Springer, Cham, pp 273–285

Marelli M, Bentivogli L, Baroni M, Bernardi R, Menini S, Zamparelli R (2014) SemEval-2014 task 1: evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), no 1, pp 1–8

Mataoui M, Hacine T, Tellache I, Bakhtouchi A, Zelmati O (2018) A new syntax-based aspect detection approach for sentiment analysis in Arabic reviews. In: 2nd international conference on natural language and speech processing (ICNLSP)

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119

Moghaddam S, Ester M (2010) Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: Proceedings of the 19th ACM international conference on information and knowledge management, pp 1825–1828

Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the seventh international workshop on semantic evaluation, pp 321–327

Montejo-Ráez A, Díaz-Galiano MC, Martínez-Santiago F, Ureña-Løpez LA (2014) Crowd explicit sentiment analysis. Knowl Based Syst 69(1):134–139

Morency L-P, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: harvesting opinions from the web. In: Proceedings of 13th international conference on multimodal interfaces—ICMI'11, pp 169–176

Nakov P, Rosenthal S, Kozareva Z, Stoyanov V, Ritter A, Wilson T (2013) SemEval-2013 task 2: sentiment analysis in Twitter. In: Joint conference on lexical and computational semantics (SEM). Volume 2: Proceedings of the international workshop on semantic evaluation (SemEval 2013), vol 2, no SemEval, pp 312–320

Napitu F, Bijaksana MA, Trisetyarso A, Heryadi Y (2017) Twitter opinion mining predicts broadband internet's customer churn rate. In: IEEE international conference on cybernetics and computational intelligence (CyberneticsCom), pp 141–146

Narr S, Hülfenhaus M, Albayrak S (2012) Language-independent twitter sentiment analysis. In: Workshop on knowledge discovery, data mining and machine learning (KDML-2012), Dortmund, Germany

Nogueira C, Santos D, Gatti M (2014) Deep convolutional neural networks for sentiment analysis of short texts. In: Proceedings of 25th international conference on computational linguistics, pp 69–78

Nozza D, Fersini E, Messina E (2016) Deep learning and ensemble methods for domain adaptation. In: IEEE 28th international conference on tools with artificial intelligence deep, pp 184–189

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics, p 271

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Empirical methods in natural language processing (EMNLP), vol 10, pp 79–86

Peñalver-Martinez I et al (2014) Feature-based opinion mining through ontologies. Expert Syst Appl 41(13):5995–6008

Peng H, Ma Y, Li Y, Cambria E (2018) Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowl Based Syst 148:55–65

Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

P'erez-Rosas V, Mihalcea R, Morency L (2013) Utterance-level multimodal sentiment analysis. In: Proceedings of the 51st annual meeting of the Association for Computational Linguistics, pp 973–982

Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S (2014) SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th international workshop on semantic evaluation, pp. 27–35

Pontiki M et al (2016) SemEval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation, pp 342–349
Poria S, Cambria E, Howard N, Bin Huang G, Hussain A (2016a) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174:50–59
Poria S, Chaturvedi I, Cambria E, Hussain A (2016b) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings-IEEE 16th international conference on data mining, ICDM, pp 439–448
Poria S, Cambria E, Gelbukh A (2016c) Aspect extraction for opinion mining with a deep convolutional neural network. Knowl Based Syst 108:42–49
Poria S, Cambria E, Bajpai R, Hussain A (2017a) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P (2017b) Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: long papers), pp 873–883
Poria S, Cambria E, Hazarika D, Mazumder N, Zadeh A, Morency LP (2017c) Multi-level multiple attentions for contextual multimodal sentiment analysis. In: Proceedings of IEEE international conference on data mining, ICDM, pp 1033–1038
Poria S, Majumder N, Hazarika D, Cambria E, Gelbukh A, Hussain A (2018) Multimodal sentiment analysis: addressing key issues and setting up the baselines. IEEE Intell Syst 33(6):17–25
Radianti J, Hiltz SR, Labaka L (2016) An overview of public concerns during the recovery period after a major earthquake: Nepal twitter analysis. In: Proceedings of the 49th annual Hawaii international conference on system sciences, pp 136–145
Ragini JR, Anand PMR, Bhaskar V (2018) Big data analytics for disaster response and recovery through sentiment analysis. Int J Inf Manag 42(May):13–24
Rain C (2013) Sentiment analysis in Amazon reviews using probabilistic machine learning. Swarthmore College
Rana TA, Cheah Y-N (2016) Aspect extraction in sentiment analysis: comparative analysis and survey. Artif Intell Rev 46(4):459–483
Rana R et al (2016) Gated recurrent unit (GRU) for emotion classification from noisy speech. arXiv preprint arXiv:1612.07778
Rani S, Kumar P (2019) A journey of Indian languages over sentiment analysis: a systematic review. Artif Intell Rev 52(2):1415–1462
Rao T, Srivastava S (2012) Analyzing stock market movements using Twitter sentiment analysis. In: ASONAM'12 Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), pp 119–123
Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications, vol 89. Elsevier, Amsterdam
Ren Y, Zhang Y, Zhang M, Ji D (2016) Context-sensitive twitter sentiment classification using neural network. In: Proceedings of the 30th conference on artificial intelligence (AAAI 2016), pp 215–221
Rosenfeld R, Fornango R (2008) The impact of economic conditions on robbery and property crime: the role of consumer sentiment. Criminology 45(4):735–769
Roy K, Kohli D, Kumar R, Sahgal R, Yu W-B (2017) Sentiment analysis of Twitter data for demonetization in India: a text mining approach. Inf Syst 18(4):9–15
Ruangkanokmas P, Achalakul T, Akkarajitsakul K (2016) Deep belief networks with feature selection for sentiment classification. In: 7th International conference on intelligent systems, modelling and simulation (ISMS), pp 9–14
Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, pp 3856–3866
Saif H, Fernandez M, He Y, Alani H (2013) Evaluation datasets for Twitter sentiment analysis A survey and a new dataset, the STS-Gold. In: Proceedings of 1st ESSEM work, Turin, Italy, vol 1096, pp 9–21
Sánchez-rada JF, Iglesias CA (2019) Social context in sentiment analysis: formal definition, overview of current trends and framework for comparison. Inf Fusion 52:344–356
Shah RR, Yu Y, Verma A, Tang S, Shaikh AD, Zimmermann R (2016) Leveraging multimodal information for event summarization and concept-level sentiment analysis. Knowl Based Syst 108:102–109
Shaikh T, Deshpande D (2016) Feature selection methods in sentiment analysis and sentiment classification of amazon product reviews. Int J Comput Trends Technol 36(4):225–230
Shi S, Zhao M, Guan J, Li Y, Huang H (2017) A hierarchical LSTM model with multiple features for sentiment analysis of sina weibo texts. In: International conference on Asian language processing (IALP), pp 379–382

Singh P, Dave A, Dar K (2017) Demonetization: sentiment and retweet analysis. In: International conference on inventive computing and informatics (ICICI 2017), pp 894–899
Singh P, Sawhney RS, Kahlon KS (2018) Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government. ICT Express 4:124
Singhal P, Bhattacharyya P (2016) Sentiment analysis and deep learning: a survey. In: Center for Indian Language Technology, Indian Institute of Technology, Bombay
Singla Z, Randhawa S, Jain S (2017) Statistical and sentiment analysis of consumer product reviews. In: 8th International conference on computing, communication and networking technologies (ICCCNT), pp 1–6
Socher R, Perelygin A, Wu J (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of 2013 conference on empirical methods in natural language processing, pp 1631–1642
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
Song K, Yao T, Ling Q, Mei T (2018) Boosting image sentiment analysis with visual attention. Neurocomputing 312:218–228
Stojanovski D, Strezoski G, Madjarov G, Dimitrovski I (2015) Twitter sentiment analysis using deep convolutional neural network. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 726–737
Stojanovski D, Strezoski G, Madjarov G, Dimitrovski I, Chorbev I (2018) Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages. Multimed Tools Appl 77(24):32213–32242
Sun X, Li C, Ren F (2016) Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features. Neurocomputing 210:227–236
Tai KS, Socher R, Manning CD (2015) Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075
Tang D, Qin B, Liu T (2015a) Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing
Tang D, Qin B, Liu T (2015b) Learning Semantic representations of users and products for document level sentiment classification. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing, vol 1, pp 1014–1023
Tay Y, Tuan LA, Hui SC (2017) Dyadic memory networks for aspect-based sentiment analysis. In: Proceedings of 2017 ACM conference on information and knowledge management—CIKM'17, pp 107–116
Trofimova TP, Pushin AN, Lys YI, Fedoseev VM (2016) Robust visual-textual sentiment analysis: when attention meets tree-structured recursive neural networks. In: Proceedings of the 2016 ACM on multimedia conference, pp 1008–1017
Twitter Statistics (2019). https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
Uysal AK, Murphey YL (2017) Sentiment classification: feature selection based approaches versus deep learning. In: IEEE international conference on computer and information technology (CIT), pp 23–30
van Hee C, Lefever E, Hoste V (2018) Exploring the fine-grained analysis and automatic detection of irony on Twitter. Lang Resour Eval 1–25
Vateekul P, Koomsubha T (2016) A study of sentiment analysis using deep learning techniques on Thai Twitter data. In: 13th International joint conference on computer science and software engineering (JCSSE), pp 1–6
Verma S, Saini M, Sharan A (2018) Deep sequential model for review rating prediction. In: 10th international conference on contemporary computing, IC3 2017, vol 2018, pp 1–6
Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012a) A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In: Proceedings of the 50th annual meeting of the Association for Computational Linguistics, pp 115–120
Wang X, Gerber MS, Brown DE (2012b) Automatic crime prediction using events extracted from twitter posts
Wang Y, Huang M, Zhu X, Zhao L (2016a) Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 606–615
Wang H, Meghawat A, Morency L, Xing EP (2016b) Select-additive learning: improving generalization in multimodal sentiment analysis. arXiv preprint arXiv:1609.05244
Wang J, Fu J, Xu Y, Mei T (2016c) Beyond object recognition: visual sentiment analysis with deep coupled adjective and noun neural networks. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI-16), pp 3484–3490
Wang X, Li Y, Xu P (2018a) A hybrid BLSTM-C neural network proposed for chinese text classification. In: IEEE sixth international conference on advanced cloud and big data (CBD), pp 311–315
Wang Y, Sun A, Han J, Liu Y, Zhu X (2018b) Sentiment analysis by capsules. In: Proceedings of the 2018 world wide web conference, pp 1165–1174

Wang J, Peng B, Zhang X (2018c) Using a stacked residual LSTM model for sentiment intensity prediction. Neurocomputing 322:93–101
Wang Y, Sun A, Huang M, Zhu X (2019) Aspect-level sentiment analysis using AS-capsules. In: The world wide web conference. ACM, pp 2033–2044
Whitehead M, Yaeger L (2009) Building a general purpose cross-domain sentiment mining model. In: WRI world congress on computer science and information engineering, CSIE, vol 4, pp 472–476
Wu D, Chi M (2017) Long short-term memory with quadratic connections in recursive neural networks for representing compositional semantics. IEEE Access 5:16077
Wu D, Cui Y (2018) Disaster early warning and damage assessment analysis using social media data and geo-location information. Decis Support Syst 111:48
Xiong S, Wang K, Ji D, Wang B (2018a) A short text sentiment-topic model for product reviews. Neurocomputing 297:94–102
Xiong S, Lv H, Zhao W, Ji D (2018b) Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. Neurocomputing 275:2459–2466
Xu F, Kešelj V (2014) Collective sentiment mining of microblogs in 24-hour stock price movement prediction. In: 16th IEEE conference on business informatics, CBI 2014, vol 2, pp 60–67
Xu K, Liao SS, Li J, Song Y (2011) Mining comparative opinions from customer reviews for competitive Intelligence. Decis Support Syst 50(4):743–754
Xu J, Tao Y, Lin H, Zhu R, Yan Y (2017) Exploring controversy via sentiment divergences of aspects in reviews. In: IEEE pacific visualization symposium (PacificVis), pp 240–249
Yanagimoto H, Shimada M, Yoshimura A (2013) Document similarity estimation for sentiment analysis using neural network. In: IEEE/ACIS 12th international conference on computer and information science (ICIS). IEEE, pp 105–110
Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1480–1489
Yang M, Qu Q, Chen X, Guo C, Shen Y, Lei K (2018) Feature-enhanced attention network for target-dependent sentiment classification. Neurocomputing 307:91–97
Yang C, Zhang H, Jiang B, Li K (2019a) Aspect-based sentiment analysis with alternating coattention networks. Inf Process Manag 56(3):463–478
Yang M, Zhao W, Chen L, Qu Q, Zhao Z, Shen Y (2019b) Investigating the transferring capability of capsule networks for text classification. Neural Netw 118:247–261
Yelp Dataset (2014)
Yoo SY, Song JI, Jeong OR (2018) Social media contents based sentiment analysis and prediction system. Expert Syst Appl 105:102–111
You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, pp 381–388
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
Yu H, Gui L, Madaio M, Ogan A, Cassell J (2017) Temporally selective attention model for social and affective state recognition in multimedia content. In: Proceedings of the 2017 ACM on multimedia conference. ACM, pp 1743–1751
Yu L, Wang J, Lai KR, Zhang X (2018) Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 26(3):671–681
Yu J, Jiang J, Xia R (2019) Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. IEEE/ACM Trans Audio Speech Lang Process 27(1):168–177
Yuan M, Tang H, Li H (2014) Real-time keypoint recognition using restricted boltzmann machine. IEEE Trans Neural Netw Learn Syst 25(11):2119–2126
Yuan Z, Wu S, Wu F, Liu J, Huang Y (2018) Domain attention model for multi-domain sentiment classification. Knowl Based Syst 155:1–10
Zadeh A, Zellers R, Pincus E, Morency L (2016) MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intell Syst*
Zhang J, Chow C (2019) MOCA: multi-objective, collaborative, and attentive sentiment analysis. IEEE Access 7:10927–10936
Zhang Y, Wallace B (2015) A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820
Zhang Z, Ye Q, Zhang Z, Li Y (2011) Sentiment classification of internet restaurant reviews written in Cantonese. Expert Syst Appl 38(6):7674–7682

Zhang Z, Zou Y, Gan C (2017) Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing 275:1407

Zhang Y et al (2018a) A quantum-inspired multimodal sentiment analysis framework. Theor Comput Sci 752:21

Zhang Z, Wang L, Zou Y, Gan C (2018b) The optimally designed dynamic memory networks for targeted sentiment classification. Neurocomputing 309:36

Zhang B, Xu X, Yang M, Chen X, Ye AY (2018c) Cross-domain sentiment classification by capsule network with semantic rules. IEEE Access 6:58284–58294

Zhao L, Huang M, Chen H, Cheng J, Zhu X (2014) Clustering aspect-related phrases by leveraging sentiment distribution consistency. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1614–1623

Zhao W et al (2017) Weakly-supervised deep embedding for product review sentiment analysis. IEEE Trans Knowl Data Eng 4347:1–12

Zhao W, Peng H, Eger S, Cambria E, Yang M (2019) Towards scalable and reliable capsule networks for challenging NLP applications. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pp 1549–1559

Zhou K, Zeng J, Liu Y, Zou F (2018) Deep sentiment hashing for text retrieval in social CIoT. Futur Gener Comput Syst 86:362

Zvarevashe K, Olugbara OO (2018) A framework for sentiment analysis with opinion mining of hotel reviews. In: Conference on information communications technology and society (ICTAS), pp 1–4