Intelligent Systems with Applications 15 (2022) 200090

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review☆,☆☆

Faisal Alshuwaier [*,1,a], Ali Areshey [2,a], Josiah Poon [3,b]

[a] KACST 6086, Riyadh 11442, Kingdom of Saudi Arabia
[b] J12 - Computer Science Building, The University of Sydney, Sydney, Australia

ARTICLE INFO

Keywords:
Deep Learning Models
Document Level
Sentiment Analysis
Applications
Improvement
SLR

ABSTRACT

Sentiment analysis has become a highly effective research field in the natural language domain and has a large scope of real-world implementations. An existing active study concentration for sentiment analysis is the development of graininess at the document level, appearing with two featured objectives: subjectivity classification, which determines whether a document is objective or subjective and sentiment detection which defines whether or not a document has a sentiment. Deep learning approaches have featured as a chance for developing these objectives with their ability to present both syntactic and semantic characteristics of text without demands for high-level attribute engineering. In this paper, we focus to produce a systematic literature review of deep learning methods for document-based sentiment analysis to determine different features in the text. In addition, this systematic literature review presents a brief survey, evaluation, enhancement of recent developments in the field of sentiment analysis techniques and applications of documents for deep learning, starting with the Convolutional Neural Network, continues to cover the Recurrent Neural Network, including Long Short-Term Memory and Gated Repetitive Units. This review also contains the implementation and application of Recursive Neural Network, Deep Belief Network, Domain-Adversarial Network Models and Hybrid Neural Network. This work considers most of the papers published when the history of deep learning began, and specifically the sentiment analysis of the documents.

1. Introduction

The development of information technologies has permitted new goals of intelligence through human-centric meaning, and it can be as simple as a customer review or a question-and-answer forum (Rana and Cheah, 2016). The fast growth rate of such large data generates numerous sources of subjective information. Sentiment analysis has been an active study topic in natural language processing (NLP), data mining, and information retrieval. Sentiment analysis presents and monitors the comprehension of information related to public views and customer feedback in different entities and attributes, such as social networks, business perspectives, and many products or services. It also evaluates people's opinions, emotions, and attitudes toward entities and attributes that can be categorized as negative, positive, or neutral (Cambria et al., 2017). Sentiment analysis has been widely studied in multiple domains which contain a variety of features, such as n-gram features.

In their important outcome on document-based sentiment analysis (DOCSA), Zhang et al. (2018) represented that scientists have generally discussed sentiment analysis at three levels of subdivisions: document level, sentence level, and aspect level. In a more thorough analysis, sentiment classification is frequently considered a document categorization. In such a categorization, document description performs a principal role and would consider the main information carried by strings as stings in a document. So, our research aims to perform sentiment analysis at the document level, which carries an opinion on the

☆ This document is to produce a literature review of deep learning methods for document-based sentiment analysis.
☆☆ The second title footnote which is a longer text matter to fill through the whole text width and overflow into another line in the footnotes area of the first page.

* Corresponding author.
E-mail addresses: shuwaier@kacst.edu.sa (F. Alshuwaier), aareshey@kacst.edu.sa (A. Areshey), josiah.poon@sydney.edu.au (J. Poon).
[1] Academic Researcher, http://www.kacst.edu.sa
[2] Programs Developer, http://www.kacst.edu.sa
[3] Senior Lecturer, https://www.sydney.edu.au

https://doi.org/10.1016/j.iswa.2022.200090
Received 7 February 2022; Received in revised form 29 April 2022; Accepted 31 May 2022
Available online 2 June 2022

entire text as shown in Fig. 1.

The concept of DOCSA was simplified by Pang et al. (2002) which generally concentrates on finding the differences of the entire inflexible document concerning one structure. For example, allocating the customer review in a document that expresses given goods.

In recent years, particular approaches for analysis of sentiment have been presented by Stone et al. (2007), who established an integrated system for basic vocabulary analysis techniques. Previously, Pang et al. (2002) used machine learning to discover the opinion explicated in a document. These standard approaches have achieved sufficient outcomes (Costa, 2018). Later, implementing deep learning as the basis to recognize and classify the automated sentiment analysis was found to have the greatest overall impact due to the multi-layer approaches of deep learning including a sensitive process for classifying the data. Currently, deep learning is applied to many NLP areas such as text, voice, and speech, and provides accurate results. Text classification is a fundamental problem in NLP and social media (Do et al., 2019; Nguyen and Le Nguyen, 2017; Zhang et al., 2018).

Among deep learning algorithms in language modeling, the Recurrent Neural Network (RNN) uses a formal influence model with fixed-length context units, which cannot be considered to be represented in all history words used. Another learning algorithm, the Convolutional Neural Network (CNN), involves multiple layers and artificiality as it provides a standard approach to map the variable length of a given sentence into sentences of fixed-size dispersed vectors. A third deep learning algorithm is the Recursive Neural Network (RecNN), which is a type of artificial neural networks (ANNs) in which links between nodes form a directed scheme that can have different metrics. Deep Neural Network (DNN) is another type of ANN with multiple layers between two types of layers. Lastly, Deep Belief Networks (DBNs) involve different hidden layers, composed of a generative stochastic ANN. DBN has been proven to be active in the representation of different features (Ain et al., 2017).

Ancient models of deep learning examined syntactic functions, subject elements, machine learning classifiers, and statistical models to recognize documents. More lately, deep learning methods have been effectively involved in NLP, which motivates the examination of how deep learning performs when given precise functions such as DOCSA (Do et al., 2019; Schouten and Frasincar, 2016).

To our understanding, our research examines the application of deep learning models to DOCSA tasks as shown also in Fig. 1. Previous researches have managed to implement sentiment analysis in general, not at the document level, which includes detecting the differences of the whole inflexible text. So, these previous researches on deep learning have not included the fields of DOCSA tasks in-depth, even in the work of (Do et al., 2019) which presented the tasks of deep learning methods to the aspect level. Deep learning approaches in the document levels

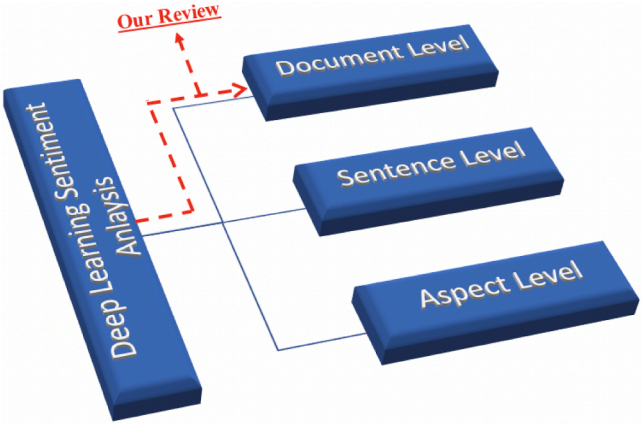

Fig. 1. Deep Learning Applications on Different Sentiment Analysis Tasks.

were also not present in detail in the work of (Habimana et al., 2019); and (Dang et al., 2020), as deep learning approaches are only covered in general terms and are not detailed as in their work. This paper focuses on presenting and comparing the most recent positive changes in neural networks in general and DOCSA in particular. This review is specially formulated for researchers in the field of text and language fields, who are seeking to explore deep neural networks as well as recent trends in DOCSA research. Finally, the current issues are highlighted in this research that needs to be addressed and make suggestions for improvement and performance including the use of new paradigms such as the use of importance of sentences taken from aggregate training data as well as the use of Clause and Discourse Connectives.

This review is conducted based on the Systematic Literature Review (SLR) which consists of several phases: ”research questions” and ”research process”, where each phase includes several steps as described in the next section.

The rest of this paper is divided as follows: Section 2 explains the methodology of this review based on SLR. Section 3 determines the assignments of DOCSA and measurement systems; Section 4 evaluates deep learning models for DOCSA, examining in detail the effects of deep learning in sentiment analysis at the document level, building design, and the efficiency of DOCSA assignments; Section 5 presents the discussion for a comparison of performance on standard datasets, opportunities, challenges and the sentiment analysis improvement in DL at the document level; The evaluation for the quality assessment of SLR is described in Section 6. Finally, the conclusion in Section 7 briefly discusses the present aspects of DOCSA.

2. Methodology

This review is conducted based on the SLR presented (Kitchenhamy et al., 2009) which this methodology consists of several phases: ”research questions” and ”research process”, where each phase includes several steps. The first phase is the research question as it involves creating a review protocol to help guide researchers through the next phases to reduce potential biases that may arise. The components of the protocol for the other phases include all elements of the review such as resources, research keywords, research expression, inclusion and exclusion criteria, quality assessment, data gathering, analysis of data and the deviations from the protocol.

2.1. Research Questions

This section presents the research questions that should be presented in this review:

- RQ1: What is the assignments of DOCSA?
- RQ2: What are the overall impact of the application of deep learning for DOCSA?
- RQ3: What is the comparison of efficacy on primary datasets?
- RQ4: What are the performances with the dataset in a different domain?
- RQ5: What are the opportunities and challenges that present in DOCSA?
- RQ6: Which impact model can improve the performance applied to the DOCSA?

2.2. Research Process

The research process has been done by manual searching for specific conference proceedings and journal papers from 2002 until 2022. The research process used in this review is as follows:

2.2.1. Resources

The research process includes journal articles and conference proceedings published between January 2000 and June 2020. The most

widely used publishers are EMNLP, IEEE Intelligent Systems, ACM, Springer, Elsevier, JAIR, AAAI, ACL and others. Furthermore, well-known databases such as Scopus, Web of Science, DBLP and Google Scholar were searched. Thus, references of all selected papers were checked to see if we had missed any relevant papers.

2.2.2. Research Keywords

The keywords used in the research questions for this review are as follows:

Deep learning, document level, sentiment analysis, CNN, RNN, RecNN, DBN, DANN, HNN, applications, improvement, opinion and enhancement.

2.2.3. Research Expressions

The procedure provided for using the research terms has been applied for this review. Keywords terms are taken from research questions related to sentiment analysis of document-level using deep learning techniques. Research expressions consist of a set of keywords categorized using the AND logical operator, and combine terms and synonyms with the OR logical operator (Kitchenhamy et al., 2009).

2.2.4. Inclusion and exclusion criteria

All peer-reviewed articles on the following topics, published between January 2002 and 2022 were included:

- Deep learning methods for document-based sentiment analysis.
- Developments in the field of sentiment analysis of documents for deep learning.
- Recently developed frameworks, standardized data sets and enhancement that are used to implement and evaluate sentiment analysis of document methods for deep learning.

We also included articles in which the literature review was only one item of the articles to those for which the literature review was the main outcome of the article. Articles related to the following topics have been excluded:

- The tasks of deep learning methods to the aspect or sentence level.
- Previous researches have managed to implement sentiment analysis in general not in the document level.

2.2.5. Quality Assessment

Each SLR was estimated for review and publication database. The criteria are based on four Quality Assessment (QA) questions:

- QA1. Is it likely that the study experiment in the literature has the English language?
- QA2: Are the inclusion and exclusion criteria for review accurate, comprehensive, and appropriate?
- QA3. Have the baseline data/studies been adequately described?
- QA4. Did the study set and extract the quality standard results from the primary study?

All questions were sorted as follows:

- QA1: Y (yes), the study experiment in the literature has English language. P (Partly), the study experiments in the literature are implicit with English and other languages; N (no), the study experiments are not done in the English language.
- QA2: Y, the inclusion criteria are fully defined in the study, P (partially), the inclusion criteria implied; N (No), the inclusion criteria are not defined and cannot be easily inferred.
- QA3: Y, information about each study is displayed; P Only brief information on primary studies is shown; N, the results of the individual preliminary studies are not defined.
- QA4: Y, the authors set quality standards results and extracted them from each primary study; P, the research question contains the quality issues addressed in the study; N, an explicit estimation of the quality of the individual primary studies has not been tested.

The scoring step are Y = 1, P = 0.5, N = 0, (information not specified). we worked to coordinate the quality assessment abstraction process as well as evaluate each paper. When there was a conflict, we exchanged opinions until we reached an agreement (Kitchenhamy et al., 2009).

2.2.6. Data Gathering

The data collected from each study are as follows:

- Scientific journals and conferences.
- Specify the type of study (SLR, Analysis, Evaluation).
- Main Topic.
- Summary of the research, including important research questions and answers.
- Quality assessment.

2.2.7. Analysis of Data

The data were presented as follows:

- Names of the researchers participating in the study as mentioned in RQ1.
- Year of publication of scientific research as mentioned in RQ1.
- The type of data used in the search as mentioned in RQ1.
- The type of domain that was used with the data as mentioned in RQ1.
- The language used in the research field as mentioned in RQ1.
- The model that was built to represent the study as mentioned in RQ2.
- Comparison of Efficacy on Primary Datasets as mentioned in RQ3.
- Representation of the accuracy/performance of the results as mentioned in RQ4.
- Performance with a specific dataset in a domain as mentioned in RQ5.
- Improvement of the performance applied to the DOCSA as mentioned in RQ6.

2.2.8. Deviations from protocol

Regarding the revision of previous unsourced papers, we have made some changes to the elements of the original protocol (Kitchenhamy et al., 2009):

- SLRs are certified as part of the aggregating evidence.
- Several of research questions are listed.
- Some information is not enough so the number of quality questions has increased.
- The data was collected and analyzed based on the research questions.
- The mean quality score was calculated to compare the study if it is improved between the years using this equation:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2}$$

3. The Assignments of Document Based Sentiment Analysis (DOCSA)

DOCSA is used primarily throughout the document as a base unit of information and it is assumed that the document is known to contain an opinion and contains opinions about a single entity. These can be reviews of the entire document and creating an opinion document.

The purpose of this study is to review all deep learning approaches to

emotional analysis using a document level that can be considered as a reference in future applied studies. We intensified our efforts on key aspects of the research, such as technical challenges and opportunities, data sets, methods proposed in each study, and areas of application.

This study identified three primary sub-tasks to DOCSA are: (I) extracting opinion words, (II) polarity detection, and (III) polarity classification, whereby the extracted opinion words are examined with the extraction of word terms, polarity detection with identification of associated words, and polarity classification with the illustration of the sentiment polarity of the words Fig. 2 Kolkur et al. (2015) shows the study design of DOCSA. This research aims to collect all the researches that were done to extract the words of opinion in the document and polarize them. It is assumed that a document has an opinion about a single object, such as a hotel, drama, or journal. This model is appropriate if a document does not consist of opinions about several objects, as is often the case in blogs and other posts. Polarity detection and classification for the whole document are done as positive, negative, or neutral. The sentences should be prepared to eliminate the pre-processing level Kolkur et al. (2015).

On the one hand of documents, Turney (2002) presented a simple learning algorithm for classifying reviews of things such as automobiles, movies, travel, and banks as recommended. The classification of a polarity review is promised by the standard semantic orientation of the sentences in the review. The algorithm scores an average accuracy of 74% when estimating more than 400 reviews. The accuracy exceeds 80% for automobile reviews and 65% for movie reviews.

Zhao et al. (2014) researched sentiment analysis of documents, and the research assumed four feature selection methods, three feature representations, and five learning methods for sentiment analysis of Chinese news comments. The experimental results showed that, excepting some feature selection methods, all were suitable for selecting features for news comments, and through a comprehensive assessment of the feature selection methods, one of them was found to be the best. Therefore, one of the learning methods outperforms other classifiers for sentiment classification.

Sharma et al. (2014) demonstrated the sorting of documents according to polarity as positive, neutral or negative opinions. Disapproval was also covered in the proposed system. Effective outcomes using document polarity provide a review of the total number of positive and negative documents shown in the system. Their proposed approach produces the sentiment words from documents and provides a summary of the total number of positive and negative documents according to the polarity detection and classification, as shown in Fig. 3 (Sharma et al., 2014).

3.1. Knowledge-based and Datasets

Santos and Gatti (2014) applied two approaches to corpora from two different domains: movies and Twitter. The movies domain used a corpus containing sentences from movie reviews, and the Twitter domain used a corpus containing Twitter messages. DOCSA is mostly used for customer reviews from e-commerce and web platforms such as Yahoo as well as newsgroups, the industrial sector, and the WebKB dataset. These are likely to be text or blog analyses, and it may be supposed that in each of these, only one structure is indicted. Therefore, other researchers randomly selected news articles from the Gigaword corpus and gathered labels to train the target sentiment classifier.

In previous studies, systems have been improved for domains such as education and movie reviews following Tan and Zhang (2008), who published research containing more than 1000 documents in different domains: education, movies, and houses. They also combined a large-scale hotel review dataset that contains many positive and negative documents.

Behdenna et al. (2018) introduced neural network models (Conv-GRNN and LSTM- GRNN) for document-level sentiment categorization. They managed document-level sentiment classification on some scale review datasets from the IMDB and Yelp Dataset Challenges.

In recent years, systems have been improved for domains such as video games, Amazon reviews, mobile reviews, product reviews, blogs, opinion reviews, and tweets (Behdenna et al., 2018).

Table 1 gives a list of publicly available datasets (Behdenna et al., 2018).

3.2. Early Approaches to DOCSA Tasks

Turney (2002) presented an earlier approach to identification. In this approach, adjectives and adverbs were produced, and then semantic analysis of produced phrases was determined using some information theory and statistics. Finally, studies were categorized as the mean semantic direction of the sentences. Although this approach was limited by the time required for computation, it was a very high achievement, and the disadvantage of processing speed will be mitigated by hardware development. The latter disadvantage might be classified by using semantic direction joint with other features in a different algorithm of classification such as supervised classification and others (Behdenna et al., 2018).

Information theory, statistics, and Latent Semantic Analysis (LSA) have also been applied to conclude the semantic direction from associations between words in the document proposed by Turney (2002).

Further expansion was achieved using the semantic network WordNet and a set of tagged words extracted from a dictionary and lexicon to determine lexical links between adjectives. A limitation of the study is the size of the corpora desired for better performance. A large corpus of text inside a given document requires significant processing size and time.

A semi-supervised learning-based method to define words according to their semantics inside documents was proposed by Esuli and Sebastiani (2005). The basic expectation of that research was that words with similar directions tend to have similar explanations.

To assign sentiment outcomes to each featured entity in the text and

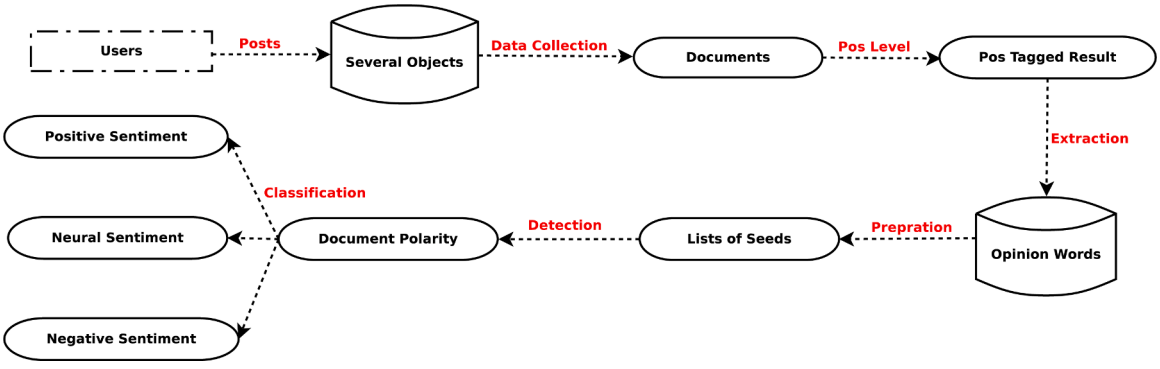

Fig. 2. Document based Sentiment Direction Model.

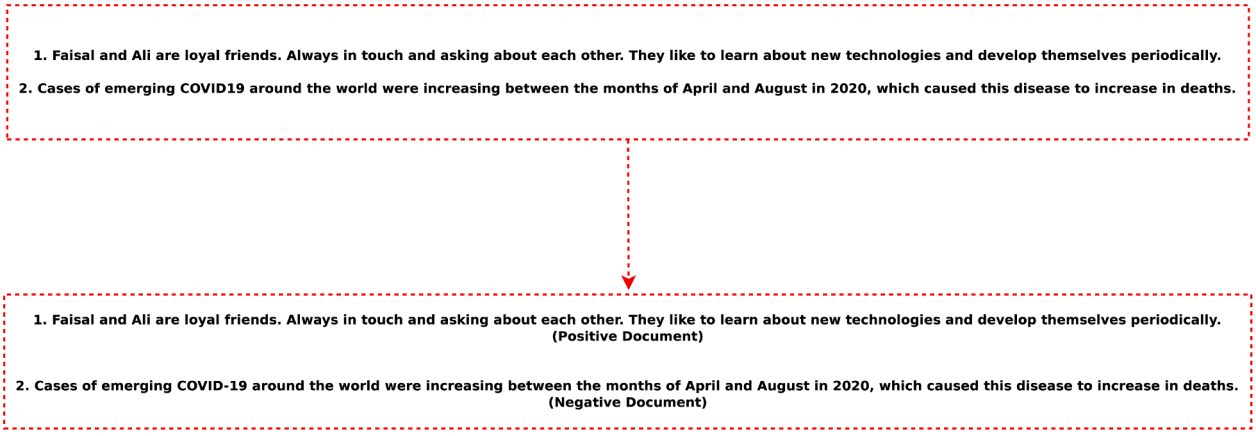

Fig. 3. Example of Document polarity-based on Extract Opinion Words.

then assign an overall subjectivity outcome to the words in the document, Godbole et al. (2007) suggested a sentiment analysis in the documents for a lexicon-based semantic approach. A lexical algorithm employing some polarity outcomes was shown in this study that explained some semantics-based algorithms that could correspond to the precision of some learning algorithms. Regarding the limitation, the geographical regions of favorable or adverse opinions for given entities have not been identified in this study.

Learning methods became familiar for sentiment analysis following productive research conducted by Pang et al. (2002). In that research, some classification algorithms were developed. Support Vector Machine (SVM) presented the best efficiency on the movie review datasets with a mean precision of approximately 80%. This research was expanded by assuring the feature chosen for Naïve Bayes (NB), resulting in classification accuracy above 85%. The limitation of this study is that the precision of NB and subjectivity similarity in the text were not included in the pre-processing step.

The similarity of subjectivity in sentiment classification was discussed using some learning methods in multi-class and regression processes with metric labeling. The outcomes determined that integration of multiple classes with other unsupervised categorization approaches resulted in improved precision. This multi-class-based approach was expanded in another study to propose objections in the context of the text in the document. However, modeling of the relationships and agreement between authors were not presented in this study by Pang and Lee (2002).

Xia et al. (2015) proposed a novel data expansion approach called dual sentiment analysis (DSA), which addresses the polarity shift problem in sentiment classification. The basic idea of DSA is to create reversed reviews that are sentiment-opposite to the original reviews and make use of the original and reversed reviews in pairs to train a sentiment classifier and make predictions. In this study, the authors focused on creating reversed reviews to assist in supervised sentiment classification.

Rong et al. (2014) presented a feature extraction capability of an auto-encoder deep learning-based method to address the dimensionality in a document. This method can offer the researchers moreover perception into capturing oriented resolution for sentiment analysis. This deep learning prediction model has been confirmed to be useful in overcoming some disadvantages while minimizing the abstraction error of the model remains an important challenge.

Morency et al. (2011) addressed the task of several different modes of sentiment analysis and control through evidence-of-abstraction experiments. The experiments explain that a learning model which integrates different domain features can be effectively used to identify sentiment in video channels via the Internet. This study made three important points. Firstly, it classified the task of several different modes of sentiment analysis and showed sentiment analysis to be functional research that can support the learning model of different domain modes, especially in text documents. Second, it determined a subset of some audio and visual features in sentiment analysis and presented recommendations on how to distinguish positive and negative classifications. Finally, it launched a new dataset online consisting of real data. This study is not sufficient for a significantly larger scale, so there is a problem with sentiment analysis at the utterance level.

He et al. (2012) presented a dynamic joint sentiment-topic model (dJST) that permits the tracking and identification of views of previous and current interest and then conducts a sentiment analysis. Both subject and sentiment dynamics are occupied by assuming that the present sentiment-topic specific word allocations are produced according to the word distributions in the last analysis (Do et al., 2019).

Zhao et al. (2017) showed more types of syntactic representation for potential semantic features. Relation embedding is applied to classify the potential semantics between targets and their identical polarity words in the documents, and subtree embedding is used to investigate the large syntactic information for each word in a document. To combine the two types of syntactic representations, a deep learning method is structured. A RecNN was used to model the subtree embedding, and then the word and subtree embedding are integrated as the improved word representation in the document for each word in the syntactic approach. Finally, a CNN was also adopted to combine the two types of syntactic illustrations to extract sentiment analysis in the datasets.

3.3. Performance Evaluation Measures of DOCSA Tasks

The research studies on semantic assessment enhanced the evolution of document-level sentiment analysis. Do et al. (2019) presented a system for evaluation and datasets for all contributors. For the measurement of the overall accuracy (OA), F-score (F1) and Accuracy (A%) can be represented by ratios, while Precision (P) and Recall (R) can be represented by ratios of entries from the matrix.

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2PR}{P + R} \tag{5}$$

$$A\% = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

TP: True Positive. TN: True Negative. FP: False Positive. FN: False

Table 1
A List of Publicly available Datasets.

No.	Dataset & Author	Domain & Language	Format	Example
1	Movie Review Dataset. (Moraes et al., 2013).	Movie, GPS, Books and Camera. (English Language).	Text Format.	The number of positive reviews is substantially greater.
2	Stanford Sentiment Tree- bank (SSTb). Santos and Gatti (2014).	Movie Reviews (8544 sentences) and Twitter (80K Sentences). (English Language).	Text format, Hyper-parameter values for the two labeled datasets.	The Positive / Negative column presents prediction results.
3	Standard Movie Dataset. (Dhande and Patnaik, 2014).	Movie Review Dataset, 2000 text files. (English Language).	Neural Network Classifier + Naïve Bayes Neural Classifier	[-1,0,1] output as 1 means expected and predicted output as a -1 and expected output as a 1 of given example.
4	IMDB and Yelp Dataset Challenge. (Tang et al., 2015).	One movie review as number of docs [348,415] and three restaurant review as number of docs [335,018, 1,125,457, 1,569,264]. (English Language).	Text Features.	MSE to measure the divergences between predicted sentiment labels and ground truth sentiment labels.
5	French Articles Dataset. (Rhanoui et al., 2019).	2003 National and international newspapers. French Articles.	Text format CNN using BiLSTM Classification.	1247 neutral tone articles, 474 positive articles and 282 negative ones.
6	IMDB, Yelp 2014 and Yelp 2015 datasets. (Raoa et al., 2018).	One Movie Review and Two Restaurant Review. (English Language).	Text Format which a softmax added layer to output the probability of classifying the document as positive, negative or neutral.	All matrices with sampling from uniform distribution in [- 0.1, 0.1].
7	IMDb Dataset, ripadvisor.com) and Yatra (yatra.com) Dataset. (Sharma and Dey, 2012).	Movie reviews + Hotel Review. (2000 Reviews). (English Language).	Text documents. ANN is a popular classifier that can be adopted for linear and non-linear text categorization.	Java based implementations on Microsoft Windows platform were used to implement the BPANN.
8	Stanford sentiment treebank dataset and IMDB dataset. (Le and Mikolov, 2014).	Movie review site and Rotten Tomatoes. 98544 sentences for training, 2210 sentences for test and 1101 sentences for validation). (English Language).	Text Format. 5-way fine-grained classification or a 2-way coarse-grained classification.	Every sentence in the dataset has a label which goes from very negative to very positive in the scale from 0.0 to 1.0.
9	The data are available from the website cs.jhu.edu/ (Glorot et al., 2011).	Large Amazon dataset consists of 22 domains (340,000 reviews). (English Language).	Text Format, Stacked Denoising Auto-encoder (SDA).	More domains with differ- ent and larger sizes, different ratios between positive and negative examples.
10	IMDB Dataset. (Zhai and Zhang, 2016).	The first one is from Movie review and the rest five datasets are all collected from Amazon. (English Language).	Text format. Filters learned by SBDAE and DAE (Denoising Autoencoder).	The top 5 most activated and deactivated words of the first 8 filters.
11	IMDB, Amazon and RCV1. (Johnson and Zhang, 2015).	Movie reviews, Electronic product reviews and Reuters news articles. (English Language).	Text format. seq-CNN.	The proximity of the embedded vectors tends to reflect the proximity in terms of the relations to the target classes.
12	IMDB, Yelp 2014 and Yelp 2013. (JChen et al., 2016).	Online Review, Yelp 2014 and Yelp2013. (English Language).	Text Format. User Product Attention (UPA) based Neural Sentiment Classification (NSC) model via the Hierarchical Long Short-term Memory (HLSTM).	UPNN achieves 3% improvement and their proposed NSC+UPA obtains 9% improvement in accuracy.
13	IMDB, Yelp and Yelp 2013 datasets. (Dou, 2017).	User and Products reviews for IMDB, Yelp 2014 and Yelp 2013. (English Language).	Text Format. Deep memory network to capture the user and product information and solve the tasks in sentiment classification at document-level.	The authors tried to minimize the cross entropy error of sentiment classification in a supervised manner.
14	Yelp 2013, Yelp 2014 and IMDB Datasets. (Xu et al., 2016).	Yelp 2013, Yelp 2014, Restaurant Reviews and IMDB for Movie Reviews. (English Language).	Text Format. Cached Long Short-Term Memory neural networks (CLSTM) which outperforms the state-of-the-art models on three publicly available document-level sentiment analysis datasets.	Yelp 2013 and Yelp 2014: The sentiment polarity of each review is 1 star to 5 stars. IMDB: popular movie review dataset ranging from 1 to 10.
15	Yelp reviews, IMDB reviews, Yahoo answers and Amazon reviews. (Yang et al., 2016).	Yelp 2013, Yelp 2014, Yelp 2015 reviews, IMDB reviews, Yahoo answers and Amazon reviews.	Text Format. GRU - based sequence encoder.	The word good may appear in a review that has the lowest rating either because users are only happy with part of the product/service.
16	Dataset from the cross-language sentiment classification evaluation of NLP and CC 2013. (Zhou et al., 2016b).	Book, DVD and Music. The dataset contains reviews in three domains including book, DVD and music. (Using English as the source language and Chinese as the target language).	Text Format. They propose an attention based LSTM network for cross-language sentiment classification.	The sentiment label vector y = (1, 0) represents the positive sentiment and y = (0, 1) represents the negative sentiment.
17	Amazon review dataset. (Li et al., 2017).	Books, DVD, Electronics and Kitchen. There are 2000 labeled reviews for each domain. (English Language).	Text Format. It is a deep models, called Adversarial Memory Network (AMN) for cross-domain sentiment classification.	Label 1 denotes positive sentiment and label 0 denotes negative sentiment.
18	ChnSentiCorp, Chn-SentiCorpEdu, ChnSentiCorpMov, and ChnSentiCorp-Hou. (Tan and Zhang, 2008).	Education, Movie, and House. There are 507 education-related documents, 266 movie-related documents and 248 house-related documents. (Chinese Language).	Four feature selection methods and five learning methods.	The total positive documents amount to 458; while the total negative documents amount to 563.

Negative.

Precision calculates the percentage of labels correctly assigned by the system. Recall calculates the percentage of labels found by the system. Accuracy and F-score represent true results (TF and TN). For subjectivity classification, sentiment detection, and classification tasks, the F-score is frequently used as the tasks are identical to information retrieval and, to estimate sentiment polarity, accuracy is used (Do et al., 2019). Moreover, For sentiment intent extraction and text class detection in document assignments, the F score is used mainly because the tasks are repetitive to retrieve information, and for assessing service points, accuracy is applied.

4. The Implementation of Deep Learning for DOCSA

Deep learning is a method of machine learning which refers to the description of data through algorithms separately from feature extraction, which is instead performed during the training process. The use of deep learning methods has achieved remarkable results for sentiment

analysis (Do et al., 2019; Nguyen and Le Nguyen, 2017). Implementing deep learning as the basis for recognizing and classifying the automated sentiment analysis provides the greatest overall impact because multi-layered deep learning approaches can have a sensitive process to classify the data (Zhang et al., 2018). Deep learning can provide accurate results when applied to a variety of NLP areas such as text, voice, and speech. Text classification is a fundamental problem in NLPs. Existing research has focused on rules based on features; however, only a limited number of studies have effectively represented the learning qualifications of deep learning approaches.

4.1. DL with Several Processing Layers

DL architectures and methods have already been developed for different layers in the NLP. Following this orientation, the current NLP study is now progressively localizing on the usage of new deep learning methods. For decades, different learning techniques aiming at NLP issues have been founded on smaller deep models (e.g., SVM and logistic regression) based on different features with a high dimensional space (Young et al., 2018). Statistical NLP has grown to handle the complex layout of natural language details. However, in its beginning, a drawback of dimensionality resulted from learning the combined probability functions of language models. This guided to the representation of learning distributed vectors embedding. This embedding directed the distributional possibility, according to which vectors with similar meanings tend to occur in a similar context. The main advantage of embedding is that the words can be captured by similarity in the context. Word embedding is often used as the initial layer in the processing step in a deep learning method (Do et al., 2019; Young et al., 2018). Bengio et al. (2003) presented a model that used a neural learning model to distribute word representations. These word embeddings were collected into sentence representations using the combined probability of word sequences and achieved the semantic neighboring of sentences in the document. Therefore, the neural language model has three components: the table look-up for scaling up the size of the neural language model, a series of interconnected words for regulating the output of the neural model, and softmax classification for word embedding to predict a multinomial probability distribution, as shown in Figure 4 (Young et al., 2018).

In addition, these word embeddings were transformed by continuous bag-of-words (CBOW), where only one word is considered in the context (Do et al., 2019; Young et al., 2018). Therefore, CBOW has three layers for NLP, namely, the input layer which contains the number of windows and vocabularies, the hidden layer which is the average of one input vector that uses the weights of W, and the output layer is to apply the soft-max function, as shown in Figure 5 (Karani, 2018).

Each word from the lexicon is finally explained as two learned vectors v_c and v_w, corresponding to context and target word embeddings, respectively. Thus, the k^{th} word in the lexicon will have

$$v_c = W_{(kc)}; v_w = W_{wk} \tag{7}$$

Generally, for any word w_i with given context word c as input,

$$p\left(\frac{w_i}{c}\right) = y_i = \frac{e^{ui}}{\sum_{i=1}^{V} e^{ui}} where v_{wi}^{T}.v_c \tag{8}$$

The second feature - hidden layers - can be structured in many forms. Each hidden layer consists of various functions stacked together to calculate nonlinear outcomes (LeCun et al., 2015). Basically, the top

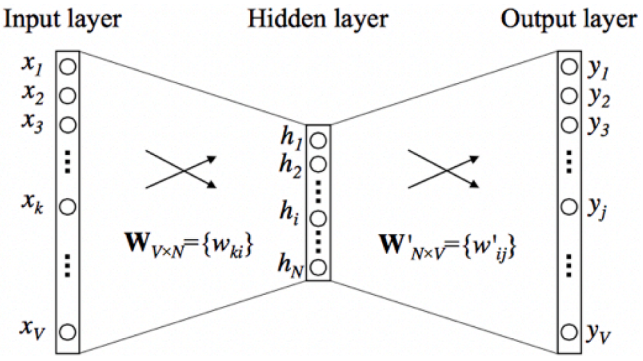

Fig. 5. Model for CBOW.

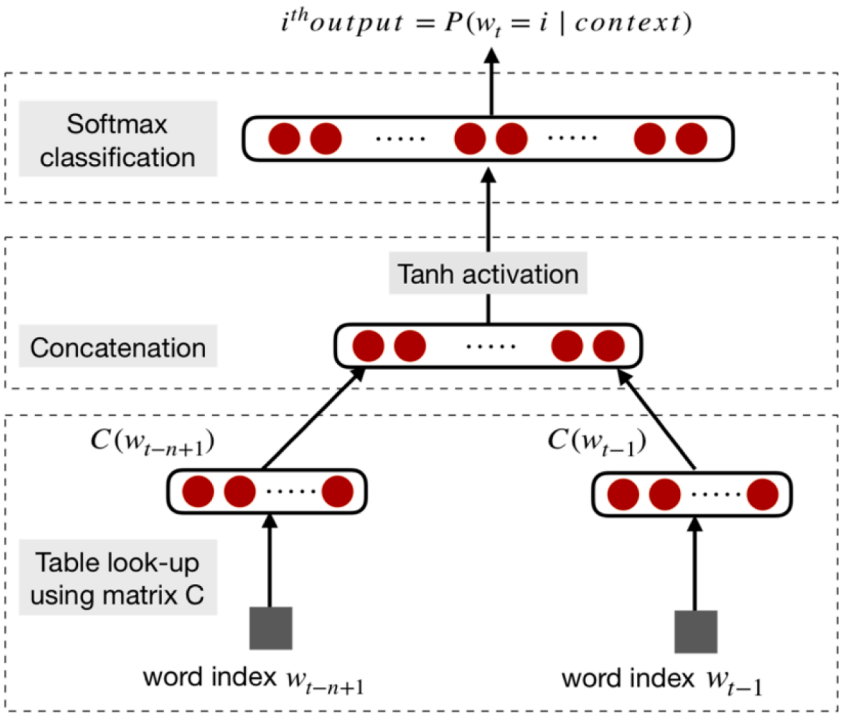

Fig. 4. Neural Language Model.

layers improved through training to achieve the complicated configuration nonlinear functions of the bottom layers and, consequently, assign more extract representations than the bottom layers (Goldberg, 2016).

The third feature - output units - performs the classification probability through labels. Assuming the output layer is G and there are L labels, the probability for a label I can be applied using the softmax function as set out below:

$$y_i = softmax(G)_i = \frac{e^{Gi}}{\sum_{L=1}^{L} e^{Gl}} \tag{9}$$

To outline the above investigation, LeCun et al. (2015) mentioned that DL with multiple layers (input, hidden and output) for NLP representation can apply a new collection of learned features beyond what has been learned at the training level to apply the output of document as modeling, extraction, translation and sentiment(Do et al., 2019), as shown in Figure 6.

4.2. Representation of Input Layer

The input layer consists of a distributed representation, or dense embedding, which represents each word as a word2vec, doc2vec, etc. to encode its semantic and syntactic properties (Do et al., 2019; Tang et al., 2016).

4.2.1. Embedding Models

Bengio et al. (2006) presented the first embedding models by using a neural network language with a divided lookup table and window size. The model can look up its constant vector by assuming a word and its anterior words. Therefore, this model led to feeding the vector to predict the probability function for its adjacent word. Mikolov et al. (2013) improved dense embedding with two different neural network models for training on a large dataset: a CBOW that achieves a word subject from an alternate word context, and the skip-gram model that predicts the context of word embedding from adjacent words of the input word, as shown in Figure 7 (Kulshrestha, 2019);(TensorFlow, 2022). The context words for each of the 8 words of this sentence in the document are defined by the window size. This figure illustrates the table of skip-grams for target words based on different window sizes 2 or 3.

Furthermore, a variety of software has been presented for training word embeddings including GloVe based on Wikipedia or Twitter, SENNA, word2vec, CBOW based on Wikipedia, Google News, or Amazon, Gigaword5, and Learning word vectors for many languages.

4.2.2. Vector Feature for DNN

DL usually depends on some information for positional or parser features, but seldom on language input (Kulshrestha, 2019). However, to produce more outstanding efficiency, several feature vectors have been supported in the DNN, together with dense embedding.

4.2.3. Grammatical Tagging - POS

There is a high prospect that the document terms are nouns that detect the importance of POS features in sentiment detection and classification, as shown in Hu and Liu (2004). The number of categorizations for POS tagging changes according to different tags, or even some tags based on Penn Tagger. In some studies, the number of POS types are nouns, adjectives, verbs, and adverbs, and some classes of chunks are NP (noun phrase), VP (verb phrase), PP (prepositional phrase), ADJP (adjective phrase), and ADVP (adverb phrase). Another feature proposed by Jebbara and Cimiano (2017) and Ma et al. (2018) to develop both sentiment detection and classification is general-sense knowledge through SenticNet and tagging (Do et al., 2019).

4.3. Prediction Task of DLs

Goldberg (2017) described a neural model to train the slope of all parameters that are calculated backward and improved with random slope extraction. Let $x = x_1, x_2, ..., x_n$ be the input, and $y = y_1, y_2, ..., y_n$ be the output from the learning algorithm with the existing labels be y∧ = y∧1, y∧2,..., y∧n. The target of the algorithm is to determine a function y = f(x) that equals the inputs with their correct label. The loss function for the whole sample is computed with an estimate of the parameter θ as the average loss (Do et al., 2019):

$$\iota(\theta) = \frac{1}{n}\sum_{i=1}^{n} \iota(f(x_i;\theta), y_i) \tag{10}$$

The most extensively used classifiers in recent studies contain some vector machine models and conditional random fields (CRF) classifiers, with examples in DOCSA tasks as CRF. All these are special models which were examined by Goldberg (2017), who studied the most adequate features of the input to predict the output and are trained with other loss functions. CRF assumes that the output y is linked through secondary annotation in a secondary graph. In the study of Chen et al. (2017), the CRF clarifies the outcome of a given label sequence as a provisional that is symmetrical to the input sequence (Do et al., 2019).

$$score_{CRF}(x, y) = p(y|x) = \frac{1}{Z_x}\Pi_{s\in S(x,y)}\phi_s(y_s, x_s) \tag{11}$$

4.4. Convolutional Neural Network Model (CNN)

This model tests and explains the adaptation of neural networks (Toyer et al., 2020), which is known for its performance in opinion analysis. The most powerful point of this model is that it assigns the maximum total of information extracted from documents using different layers.

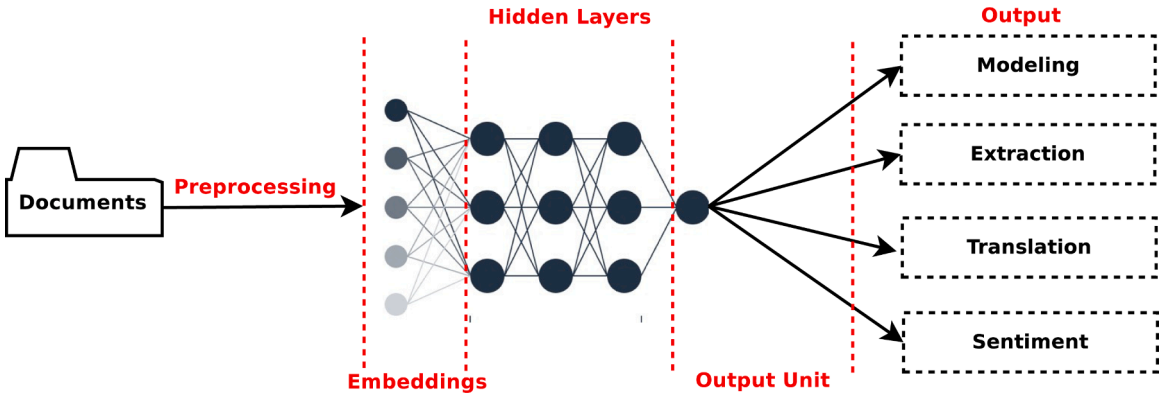

Fig. 6. DL for NLP.

Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot
3	[The **wide** road shimmered in] the hot sun.	wide, the wide, road wide, shimmered wide, in
	[The wide road **shimmered** in the hot] sun.	shimmered, the shimmered, wide shimmered, road shimmered, in shimmered, the shimmered, hot
	The wide road shimmered [in the hot **sun**].	sun, in sun, the sun, hot

Fig. 7. 6 Skip-Gram Model by Defining a Window Size.

4.4.1. Structure

The study of Rhanoui et al. (2019) proposed the following structure composed of convolution parts, which are described in more detail below in Figure 8 (Rhanoui et al., 2019); (Ghosh, 2022) and the points.

- **Pre-processing part**: In this stage, data cleansing and pre-processing are implemented to represent the distributed document using Doc2Vec embedding to prepare data for convolution. The resulting vector is passed as an input to the next layer.
- **Convolution part**: In this stage, max pooling and convolution layers are determined for feature extraction to extract high-level features from the document. The output of this layer is the input of the next stage.

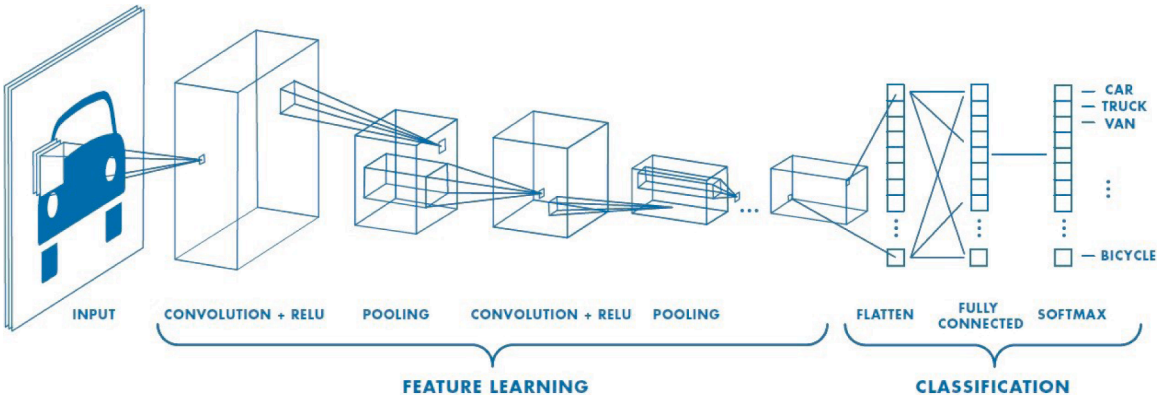

Fig. 8. Structure of CNN.

- **Final Part**: Fully connected layers with softmax and dropout are implemented for document sentiment classification. The output of this stage is the final extraction classification of the document, such as positive, negative, or neutral.

4.4.2. Application of CNN Model in DOCSA Tasks

The encouragement for using the CNN model in DOCSA tasks is the impression that the objective in the document may contain the word term and determine extraction polarity, regardless of their location. The CNN is efficient for learning to discover those features with its structure in Fig. 8 and is thus able to extract internal samples from data regardless of their position. This is adequate for recognizing fixed-length terms (Do et al., 2019; Goldberg, 2017). In addition, there is another interest in which the CNN is a non-linear model to adapt the data than linear models for example the CRF which does not request large features such as fixed language principle (Poria et al., 2016). The CNN model in DOCSA tasks has been used in many review domains, as shown in Table 2.

CNN-based models have been utilized to recognize document-level sentiment analysis by processing many filters to examine local features from the input sentence in the document.

Johnson and Zhang (2017) proposed a low-complexity word-level deep CNN structure for document level that can effectively outline long-range guides in text. The authors closely evaluated the deepening of word-level CNNs to get global representations of text and found a simple network architecture with which the best performance can be achieved by adding the network depth without increasing the computational cost by using a deep pyramid CNN.

Conneau et al. (2017) showed a new architecture using a Very Deep Convolutional Neural Network (VDCNN) for the document level which uses only small convolutions and pooling operations. The authors also presented that the performance of this architecture increases with depth: using more than 25 convolutional layers, they also announced developments over the state-of-the-art on several public text classification tasks in the document.

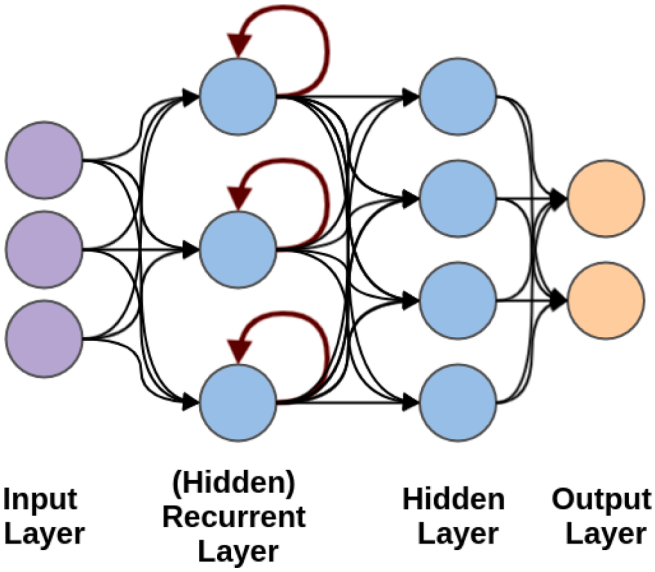

Fig. 9. The Architecture of RNN.

Other studies have utilized double or multi-tasking CNNs, showing that CNN can give other advantages. Examples of multiple CNNs can be found in Yoon and Kim (2017); Johnson and Zhang (2017); Bongirwar (2015).

Table 2
Application of CNN Model in DOCSA Tasks.

Study	Research	Dataset	Domain	Language	Model	Performance
		AG	News			Acc: 93.13%
		Sogou	News			Acc: 98.16%
		Dbpedia	Ontlogy			Acc: 99.12%
		Yelp.p	Reviews			Acc: 97.36%
S1	(Johnson and Zhang, 2017).	Yelp.f	Reviews	English + Romanized Chinese	Unsupervised Embedding + DPCNN	Acc: 64.20%
		Yahoo	Yahoo			Acc: 76.10%
		Ama.f	Website			Acc: 65.19%
		Ama.p	Reviews			Acc: 96.68%
			Reviews			
		AG	News News Ontlogy	English + Romanized Chinese		Acc: 91.33%
		Sogou	Reviews			Acc: 96.82%
		Dbpedia	Reviews			Acc: 96.82%
		Yelp.p	Yahoo			Acc: 98.71%
S2	(Conneau et al., 2017).	Yelp.f	Website		VDCNN	Acc: 95.72%
		Yahoo	Reviews			Acc: 64.72%
		Ama.f	Reviews			Acc: 73.43%
		Ama.p				Acc: 63.00%
						Acc: 95.72%
		IMDB	Movie	English		
		Amazon	Reviews			Acc: 92.33%
S3	(Johnson and Zhang, 2015).	RCV1	Electronic		BoW-CNN and Seq-CNN	Acc: 92.86%
			Prod.			Acc: 90.67%
			Reuters News			
S4		French	Newspapers	French		
	(Rhanoui et al., 2019).	Articles			CNN-BiLSTM with Doc2vec	Acc: 90.66%
		Dataset				
S5	(Yoon and Kim, 2017).	SemEval Task 4	Twitter	English	MutiChannel CNN-BiLSTM	Acc: 51.9-70%
		IMDB	Movie Rev.	English		Acc: 45.30%
S6	(Tang et al., 2015).	Yelp2013	Restaurant Re.		Conv-GRNN	Acc: 65.10%
		Yelp2014	Restaurant Re.			Acc: 67.10%
		Yelp2015	Restaurant Re.			Acc: 67.60%
S7	(Gu et al., 2017).	PM Amazon	Smartphone	English	Single	Acc: 84.87%
		PM Taobao	Shirt	Chinese	CNN	Acc: 98.26%

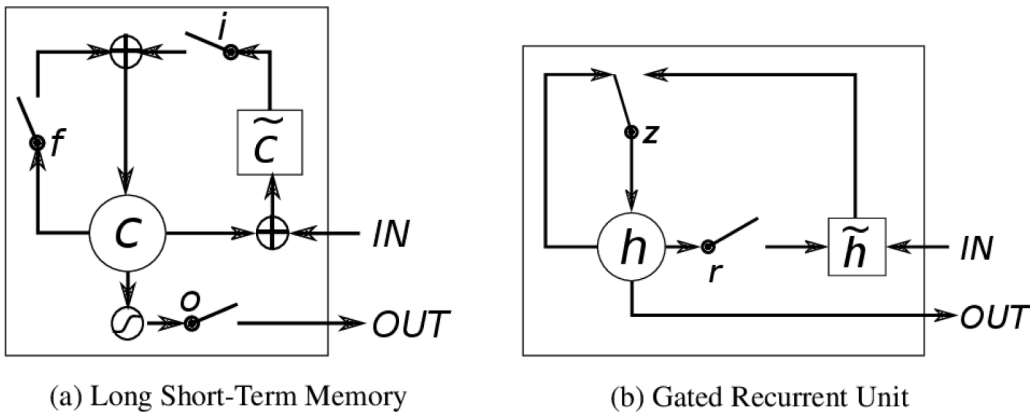

Fig. 10. The Demonstration between LSTM and GRU.

4.5. Recurrent Neural Network Models (RNN)

RNNs have become common in sentiment analysis tasks. The objective of RNN models is that a limited-size vector performs as a document or sentence by providing each token with a recurrent unit, so it can occupy the essential consecutive language nature (Do et al., 2019; Goldberg, 2016). Based on a comparison between CNN and RNN models, RNN has adaptable counting levels where the final result from RNN is subject to the previous counting, making it adequate for occupying context dependencies in language as well as for modeling with different text lengths (Tang et al., 2016).

4.5.1. Counting of RNN Models

The common RNN model establishes a model with direct periods in their implicit connection to the network (Do et al., 2019; Elman, 1991). This model shows where the implicit case is subject to the input and past implicit case, with the same function and the same set of parameters being used at every time level (Goldberg, 2016). The idea of this model is to use the recursive learning of phrase-level sentiments for each expression and provide that to longer documents the way humans

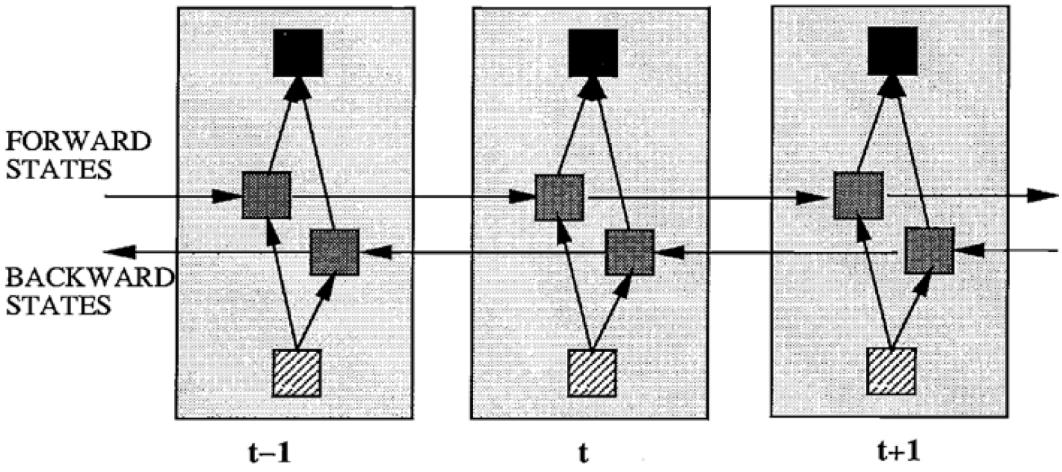

Fig. 11. The Structure of BRNN.

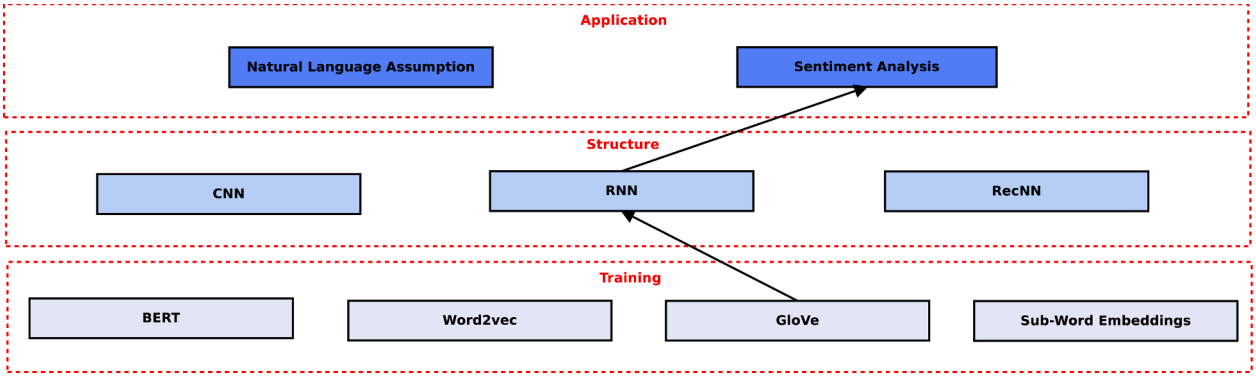

Fig. 12. BRNN Model in DOCSA.

Table 3
Application of RNN Model in DOCSA Tasks.

Study	Research	Dataset	Domain	Language	Model	Performance
S8	(Tang et al., 2015).	IMDB	Movie	English	LSTM-GRNN	Acc: 45.30%
		Yelp2013	Reviews			Acc: 65.10%
		Yelp2014	Restaurant Re.			Acc: 67.10%
		Yelp2015	Restaurant Re.			Acc: 67.60%
			Restaurant Re.			
S9	(Yang et al., 2016).	Yelp2013	Yelp Review	English	HAN	Acc: 68.20%
		Yelp2014	Yelp Review			Acc: 70.50%
		Yelp2015	Yelp Review			Acc: 71.00%
		IMDB	IMDB Review Soci., Cul., Sc.			Acc: 49.40%
		Yahoo	Amazon Rev.			Acc: 75.80%
		Amazon				Acc: 63.60%
S10	(JChen et al., 2016).	IMDB	Movie Reviews	English	NSC+UPA	Acc: 53.30%
		Yelp2013	Restaurant Rev.			Acc: 65.00%
		Yelp2014	Restaurant Rev.			Acc: 66.70%
S11	(Dou, 2017).	IMDB	Tang (2015)	English	UPDMN	Acc: 46.50%
		Yelp2013	Yelp Website			Acc: 63.90%
		Yelp2014	Yelp Website			Acc: 61.30%
S12	(Long et al., 2017).	IMDB	Eye-tracking	English	LSTM+CBA+UPA	Acc: 52.30%
		IMDB2	Eye-tracking	English	LSTM+CBA+LA	Acc: 90.10%
		Yelp2013	Newspapers	French	LSTM+CBA+UPA	Acc: 65.40%
		Yelp2014	Newspapers	Dutch	LSTM+CBA+UPA	Acc: 66.40%
S13	(Amplayo et al., 2018).	IMDB	Tang (2015).	English	HCSC	Acc: 54.20%
		Yelp2013	Tang (2015).			Acc: 65.70%
S14	(Wu et al., 2018).	IMDB	Tang (2015).	English	HUAPA	Acc: 55.00%
		Yelp2013	Tang (2015).			Acc: 68.30%
		Yelp2014	Tang (2015).			Acc: 68.60%
S15	(Qian et al., 2015).	Stanford	Movie reviews	English	TG-RNN	Acc: 86.30%
		Sentiment	Movie reviews		TE-RNN	Acc: 86.60%
		Treebank	Movie reviews		TE-RNTN	Acc: 87.70%
S16	(Mikolov et al., 2010).	WER	Oracle WER	English	3xRNN + RT09 LM	PER: 22.30%
S17	(Socher et al., 2013).	Pang and Lee (2005)	Movie Review	English	RNTN Pos.	Acc: 71.40%
					RNTN Neg.	Acc: 81.80%
S18	(Xu et al., 2016).	IMDB	Movie Review	English	CLSTM	Acc: 42.10%
		Yelp2013	Restaurant Rev.			Acc: 59.20%
		Yelp2014	Restaurant Rev.			Acc: 59.40%
S19	(Yin et al., 2017).	TripAdvisor	Hotel Review	English	Input encoder and LSTM	Acc: 46.65%
		BeerAdvocate	Hotel Review			Acc: 38.25%
S20	(Zhou et al., 2016b).	NLP-CC 2013	Book	Chinese	LSTM Multi-Channel	Acc: 82.20%
			DVD			Acc: 83.50%
			Music			Acc: 80.60%
S21	(Li et al., 2017).	Amazon	Books	English	Adversarial Memory Network	Acc: 84.71%
			DVD			Acc: 82.96%
			Electronics			Acc: 83.50%
			Kitchen			Acc: 90.00%
S22	(Sage et al., 2019).	Purchase Orders	ID Number Quantity	European	RNN	F1: 90.60%
				English		P: 90.70%
						R: 90.50%
						F1: 96.40%
						P: 95.50%
						R: 97.40%
S23	(Zheng et al., 2018).	Yelp2016 Amazon	Yelp Website	English	TextRNN	Acc: 44.33%
			Electronics		TextRNN	Acc 56.36%

explain languages-making sentiment opinion in one direction and one sentence at a time. It is certainly possible that expressions that occur early on in a review need to be evaluated differently from those in the last input, to evaluate their importance for sentiment classification in both recurrent and hidden layers, as shown in Fig. 9 (Timmaraju and Khanna, 2015); (Chaturvedi, 2022). The main function of RNN is that the neurons form a directed cycle, which develops response loops within the RNN. The main objective of processing sequential information on the principle of the inside memory is occupied by the managed cycles and loops (Dang et al., 2020; Wang and Liu, 2015).

The RNN model has more significant features over other forward neural networks. The original CNN used several arguments at each layer, whereas the arguments in RNN are the same at each level, which decreases the number of arguments needed for learning (Dang et al., 2020). Another feature is that the final result of one state depends on the anterior state. It can be said that RNN is superior to CNN to have forward counting memory, which makes it advanced in information flow processing.

On the other hand, the basic RNN has a considerable weakness in terms of appearance tendency issues. As explained earlier because the primary role of the tendency is to set the arguments to recover the tendency, extremes make it complex to resolve on which side to modify the arguments while discharging tendency causes an unstable learning process (Dang et al., 2020).

However, the basic RNN has restrictions caused by the tendency. It may disappear because this exists during the process, making it complex to train some inputs. This restriction has been mitigated with the establishment of networks such as long short-term memory (LSTM) and gated recurrent units (GRU) (Dang et al., 2020).

Fig. 10 illustrates the differences between LSTM and GRU (Chung et al., 2014).

The LSTM model has three structures: input, forget, and output gate. Geron (2017) presented the working of these gates as follows. The input gate determines if the new input should be allowed in the network. Then, information should be forgotten from the previous unit using the forget gate. Finally, the output gate receives the impact at the current

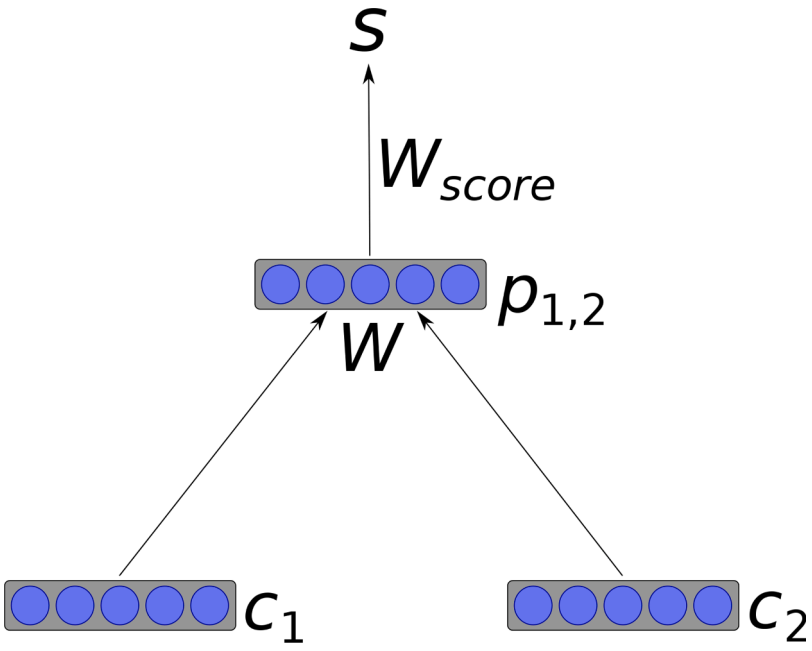

Fig. 13. RecNN Structure.

time step. On the other hand, GRUs have updated gates. The initial controls in the unit and the final gate allow the unit to start the first input from the sequence and forget the previous ones to approach the output (Geron, 2017; Sharma and Dey, 2012).

4.5.2. The Structure of Bidirectional RNN

In training, different studies create some computations based on the future words, and thus, the bidirectional RNN (BRNN) models are presented, some integrating a forward and a backward state to examine the information from preceding (t-1) and following codes (t+1) (Graves, 2008). The idea of the BRNN, as shown in Fig. 11 (Schuster and Paliwal, 1997) is to divide the state codes of a regular RNN (RRNN) into a part that is in the control of forwarding and backward states. Outcomes from forwarding states in (t+1) are not linked to inputs of backward states in (t-1), and vice versa. The BRNN can mainly be focused on with the same algorithms as an RRNN because there is no integration between the two types of state codes and, as a result, can be uncovered into a common forward state.

4.5.3. Mechanism of BRNN in DOCSA

Text classification or document is also a downstream technique of word embedding. In this technique, the pre-trained can be applied by word vectors (GloVe) and BRNN with multiple hidden layers in the structure level, as shown in Fig. 12. This model uses the vectors to define whether a text sequence in the document of unlimited length carries positive or negative emotions at the application level (Maas et al., 2011).

4.5.4. Application of RNN Model in DOCSA Tasks

RNNs and BRNNs have had a huge presence in document-level categorization in the consumer review domain, as shown in Table 3.

Different RNN-based models occupied the benefit of the BRNN to re-symbol the context. The attention method has seen the ability to support RNN to deal with ideal illustrations of a document and to occupy long-range needs at a low computational cost. In the same layer, Yang et al. (2016) planned a hierarchical awareness network (HAN) model that processes different levels of Bi-GRU with awareness to determine applicable document illustration. The HAN model structures the illustration of the code that assembles the sentence representations to provide a full document illustration sent to the classifier. In the same way, Long et al. (2017) showed a model with a cognition-based awareness model called the LSTM, which is learned by some tracking data. Originally, the LSTM+CBA model consumed the time of the sentence. Subsequently, to be applicable in the construction of document representation, the attention model structured from the predicted reading time of the document features is provided to classify the sentiment analysis and the time (Lai et al., 2015; Zhang et al., 2018).

Scientists have established that it is helpful to examine sentiment container preference and the document in sentiment analysis. JChen et al. (2016) introduced a model named neural sentiment classification and user product attention (NSC+UPA). It simultaneously learned sentences and documented performances using LSTM. Then, it requested the attention mechanism to the created representations for prioritizing the most contributing user and document details. In the same way, Dou (2017) inspired a model-styled user and document with a deep memory network, which integrated LSTM and the deep memory model. This model helped the representation of the input document. It also used LSTM to structure the document representation. Eventually, to count the ratings of each document, the model provided deep memory layers composed of the attention-based models. In the same style, Wu et al. (2018) offered a model called hierarchical user attention and product attention (HUAPA). This model independently examines the user and the document details. Then, it assembles them to make the output.

4.6. Recursive Neural Network Model (RecNN)

4.6.1. Structure

RecNN models are linguistically driven in that scientists examine syntactic structures and attempt to understand compositional semantics. Conceivably, natural language explains a natural recursive structure, placing words and phrases in a hierarchical method. Thus, tree-structured models can better construct the syntactic performance of sentence structures. Generally, in a recursive neural model, the vector representation of each node in the tree structure is studied from the performance of all its nodes using a weight matrix W, which participates across the entire model (Do et al., 2019; Socher et al., 2013).

An example process of vector representation in the RecNN is shown in Fig. 13 (Zhang et al., 2018); (Wikipedia, 2022). The vector of node "very interesting" is organized from the vectors of the node "C_1" and the node "C_2". Similarly, the node "$p_{1,2}$" is composed of the phrase node "W" and the word node "S" (Zhang et al., 2018).

Table 4
Application of RecNN Model in DOCSA Tasks.

Study	Research	Dataset	Domain	Language	Model	Performance
S24	(Al-Sallab et al., 2017).	ATB	Newswire Articles	Arabic	AROMA	Acc: 86.50%
		Tweets	Twitter			Acc: 76.90%
		QALB	Online Comments			Acc: 79.20%
S25	(Socher et al., 2011).	MR	Movie Review	English	RAE	Acc: 77.70%
		MPQA	Customer Review			Acc: 86.40%
S26	(Preethi et al., 2017).	Movie Data	Movie	English	RNN	Acc: 90.47%
		Restaurant	Restaurant Re.			Acc: 90.340%
S27	(Yuan and Zhou, 2015)	SemEval-2013	York University Website	English	One-hid.-layer	Acc: 63.71%
					Two-hid.-layer	Acc: 62.45%
					RNTN	Acc: 59.32%
					One-hid. TREE	Acc: 84.17%
					Two-hid. TREE	Acc: 80.68%
S28	(Van et al., 2017).	Standford	Amazon Review	English	RNTN	Acc: 85.40%
		Sentiment			DRNN	Acc: 86.60%
		Treebank			TE-RNTN	Acc: 87.70%

Table 5
Application of DBN Model in DOCSA Tasks.

Study	Research	Dataset	Domain	Language	Model	Performance
S29	(Zhou et al., 2016a).	Amazon Rev.	Books	EN-FR	WSDNN	Acc: 83.61%
			DVD	EN-GE		Acc: 82.40%
			Music	EN-JP		Acc: 78.26%
				FR-EN		Acc: 80.14%
				GE-EN		Acc: 80.95%
				JP-EN		Acc: 81.89%
S30	(Mikolov et al., 2013).	Article	Political	Korean	DBNWV	Acc: 81.80%
S31	(Ruangkanokmas et al., 2016)	Movie Rev.	Movie	English	DBNFS	Acc: 72.20%
			Kitchen			Acc: 75.20%
			Ele.			Acc: 75.00%
			Books			Acc: 66.00%
			DVD			Acc: 69.60 %

4.6.2. Application of RecNN Model in DOCSA Tasks

Al-Sallab et al. (2017) presented a recursive deep learning model for sentiment mining in Arabic. It was demonstrated by addressing several disadvantages that arise when providing the recursive autoencoder model to classify sentiment mining in Arabic. Arabic-specific challenges, including morphological complexity and language multiplicity, were classified by modeling the semantic structure at the Arabic morpheme level after performing morphological tokenization. The researchers presented to provide word sentiment embedding in the document to perform multiple features, including syntactic classification and sentiment analysis.

Socher et al. (2011) provided a sentiment analysis phrase model in a recursive neural tensor network. This model outperforms all previous methods for different measured sizes. It compresses the state of the art in individual sentence positive or negative classification up to 85.00%. The accuracy of predicting fine-grained sentiment analysis at the document

Table 6
Application of HNN Models in DOCSA Tasks.

Study	Research	Dataset	Domain	Language	Model	Performance
S32	(Liu et al., 2017).	RCV1	Texts	English	XML-CNN	Acc: 96.86%
		EUR-Lex.	EUROVOC			Acc: 76.38%
		Amazon-12K	Amazon			Acc: 95.06%
		Amazon-670K	Amazon			Acc: 35.39%
		Wiki-30K	wiki. social			Acc: 84.06%
		Wiki-500K	wiki. social			Acc: 59.85%
S33	(Ghosh et al., 2016).	Movie Rev.	Book	English	RBM+PNN	Acc: 81.00%
			DVD			Acc: 81.10%
			ELEC.			Acc: 80.10%
			KIT			Acc: 80.20%
S34	(Goebel et al., 2022).	SemEval 2016 Go.	Social	English	DBLSTM	Acc: 84.82%
			Twitter			Acc: 85.86%
S35	(Ravi and Ravi, 2016).	News	Viz. News	Hindi	RBF Network	Acc: 84.88%
		Facebook	Comments			Acc: 86.01%
S36	(Chen et al., 2016).	Live Journal	Journal	English	BPN	High Performance
		Review Center	Comments			Save Time
S37	(Vateekul and Koomsubha, 2016).	Tweets	Twitter	Thai	DCNN	Higher in Accuracy
					LSTM	

Table 7
Comparison of Efficacy on Primary Datasets.

	CNN	RNN	RecNN
Advantages	* Occupy potential syntactic and semantic features of sentimental analysis. * Obtain identical performance. * No need to require linguistic structure for the text in the document.	* Better for the text classification. * Preserves the order of the word sequence. * Ability to produce a fixed size vector.	* Learn tree structure. * Structure representations inside the document. * Simple Structure.
Disadvantages	* Smaller number of parameters.	* Examine a little pre-trained word embedding. * Insufficiency to occupy long-term dependencies.	* Require more parameters. * Lack of research in the application.

level reached 80.00%, an improvement of 9.70% over the bag-of-features baselines.

Preethi et al. (2017) examined a new method of RecNN with a deep learning system for sentiment analysis of reviews. The presented RecNN-based Deep-learning Sentiment Analysis (RDSA) showed areas that are close to the location of the user by inspecting the different reviews and calculating the result.

Yuan and Zhou (2015) also investigated the application of RecNN for sentiment analysis of tweets. The proposed RecNN based largely infused with symbols and short hands presented challenges as a sentiment analysis classification. In this project, the authors experimented with different classifications of neural nets and evaluated how models adjust the data set in which the quality of the data and model structures. These structures have experimented with a three-hidden-layer RNN and a recursive neural tensor network (RNTN).

Van et al. (2017) joined CNN and RecNNs into a new model architecture. In addition, they utilized transfer learning from a large document-level classified sentiment dataset to develop word embedding in their models. The resulting models exceed all recent and previous studies for CNN and RecNN. Table 4 provides an overview of the RecNN models.

4.7. Deep Belief Network Models (DBN)

DBNs contain various hidden layers. These networks have proven to be effective in representing characteristics for solving all issues related to categorized analyzes because they use unlabeled data Ain et al. (2017). Table 5 presents an overview of DBNs.

4.8. Domain-Adversarial Network Models (DANN)

Based on domain initialization theory, prediction is made using overlapping attributes between sources and target domains, where sentiment analysis is represented at the document level. Janine and others. Solve a model called Anti-Domain Neural Network (DANN), which is trained based on gradient-based optimization directly on the data to the classification network. The DANN model trains the classifier using both data and nested and unnamed attributes from the source and target domain, respectively. Then, as the training progresses, the labeled and named attributes from the source domain and random attributes for the change between domains are listed. Similarly, (Habimana et al., 2019) made a competing memory network (AMN) model containing sentiment and domain classification units within the document. So both modules are trained together to reduce sentiment analysis error and allow the domain classifier not to isolate domain samples. The attention mechanism is then integrated into the model to facilitate the selection of keywords, which are important for sentiment analysis and are involved between the source and target domains Habimana et al. (2019).

4.9. Hybrid Neural Network Models (HNN)

In each model, there are benefits and drawbacks to any previously presented, and several studies have attempted to present mixed solutions at document levels, such as (Ain et al., 2017; Chen et al., 2016; Ghosh et al., 2016; Goebel et al., 2022; Liu et al., 2017; Ravi and Ravi, 2016; Vateekul and Koomsubha, 2016; Zharmagambetov and Pak, 2015).

Liu et al. (2017) presented the first experience at requesting deep

Table 9
Performance with Yelp2014 dataset in Restaurant domain.

Dataset	Domain	No.	Research	Model	Accuracy
Yelp2014	Restaurant	S8	(Tang et al., 2015)	LSTM-GRNN	67.10%
		S9	(Yang et al., 2016)	HAN	70.50%

Table 10
Performance with Yelp2015 dataset in Restaurant domain.

Dataset	Domain	Study No.	Research	Model	Accuracy
Yelp2015	Restaurant	S8	(Tang et al., 2015)	LSTM-GRNN	67.60%
		S9	(Yang et al., 2016)	HAN	71.00%

Table 8
Performance with IMDB dataset in Movie Review, Eye-tracking data and Newspapers domain.

Dataset	Domain	Study No.	Research	Model	Accuracy
IMDB	Movie Review	S3	(Johnson and Zhang, 2015)	BoW-CNN and Seq-CNN	92.33%
		S6	(Tang et al., 2015)	Conv-GRNN	66.00%
		S8	(Tang et al., 2015)	LSTM-GRNN	45.30%
		S9	(Yang et al., 2016)	HAN	49.40%
		S10	(JChen et al., 2016)	NSC+UPA	53.30%
		S11	(Dou, 2017)	UPDMN	46.50%
		S13	(Amplayo et al., 2018)	HCSC	54.2%
		S16	(Wu et al., 2018)	HUAPA	55.00%
		S18	(Xu et al., 2016)	CLSTM	42.10%
	Eye-tracking data	S12	(Long et al., 2017)	LSTM+CBA+UPA	52.30%
	Newspapers	S12	(Long et al., 2017)	LSTM+CBA+LA	90.10%

Table 11
Performance with Yelp2013 dataset in Restaurant and Yelp Website domain.

Dataset	Domain	Study No.	Research	Model	Accuracy
Yelp2013	Restaurant	S6	(Tang et al., 2015)	LSTM-GRNN	65.10%
		S10	(JChen et al., 2016)	NSC+UPA	65.00%
		S18	(Xu et al., 2016)	CLSTM	59.20%
	Yelp Website	S11	(Dou, 2017)	UPDMN	63.90%
		S12	(Long et al., 2017)	LSTM+CBA+UPA	65.40%
		S13	(Amplayo et al., 2018)	HCSC	65.70%
		S14	(Wu et al., 2018)	HUAPA	68.30%

Table 12
Performance with Yahoo dataset in Yahoo Website domain.

Dataset	Domain	Study No.	Research	Model	Accuracy
Yahoo	Yahoo Website	S1	(Johnson and Zhang, 2015)	DPCNN	69.42%
		S2	(Conneau et al., 2017)	VDCNN	64.72%
		S9	(Yang et al., 2016)	HAN	75.80%

Table 13
Performance with Amazon dataset in Electronic Product and Amazon Review domain.

Dataset	Domain	Study No.	Research	Model	Accuracy
Amazon	Electronic Product	S3	(Johnson and Zhang, 2015)	BoW-CNN and Seq-CNN	92.33%
		S7	(Gu et al., 2017)	Single CNN	84.87%
		S23	(Zheng et al., 2018)	TextRNN	56.36%
	Amazon Review	S9	(Yang et al., 2016)	HAN	63.60%
		S28	(Van et al., 2017)	RNTN	85.40%
				DRNN	86.60%
				TE-RNTN	87.70%
	Amazon-12K	S32	(Liu et al., 2017)	XML-CNN	95.06%
	Amazon-670K			XML-CNN	35.39%

Table 14
Performance with RCV1 dataset in Reuters News domain.

Dataset	Domain	Study No.	Research	Model	Accuracy
RCV1	Reuters News	S3	(Johnson and Zhang, 2015)	BoW-CNN and Seq-CNN	96.86%
		S32	(Liu et al., 2017)	XML-CNN	84.87%

learning to Extreme Multi-label Text Classification (XMTC), with a CNN model that is customized for multi-label classification in detail. With a comparison evaluation of many state-of-the-art methods on different standard datasets where the number of labels is up to 670,000, the authors showed that the presented model effectively scaled to the largest datasets, and regularly created the best results on all the datasets. Table 6 provides an overview of HNNs.

5. Discussion

In this section, we will try to answer all the research questions, present all the data analyzed for this study, compare the efficacy of the primary dataset, summarize the opportunities and challenges of SLR that are present in DOCSA and provide improvements in sentiment analysis from the platform for documents of SLR.

5.1. Comparison of Efficacy on Primary Datasets

The previous study has given awareness of the different methods and models selected by studies for language processing assignments. The results depend not only on the selected model and configuration but also on the semantics to be resolved. There is a high prospect that the document terms are nouns that detect the importance of POS features in sentiment detection and classification, as shown in Hu and Liu (2004). This monitoring is supported by a comparison of the models. CNN models can be highly successful. This is adequate for recognizing fixed-length terms. In addition, CNN is a non-linear model and can thus fit the data better than linear models such as the CRF, and does not request large features such as a fixed language principle (Goldberg, 2017; Poria et al., 2016). Furthermore, CNN models, as presented by Johnson and Zhang (2017), have proposed a low-complexity word-level CNN architecture for the document level that can effectively outline long-range guides in text. As pointed out by Tang et al. (2016), this influences the application of CNNs to languages with morphologically-rich texts such as Romanised Chinese and French. Such cases benefit from a model capable of understanding long-term dependencies, such as RNN or RecNN (Do et al., 2019).

RNNs are tremendous because they join two features: (1) distributed hidden states that recognize them to effectively save information from previous calculations; and (ii) nonlinear dynamics that are most appropriate for the nonlinear nature of data. Based on a comparison between CNN and RNN models, RNN has adaptable counting levels that the final result from RNN is subject to the previous counting, making it adequate for occupying context dependencies in language as well as adequate to model different text lengths (Tang et al., 2016). The main function of RNN is that the neurons form a directed cycle, which develops response loops within the RNN. The main objective of processing sequential information on the principle of the inside memory occupied

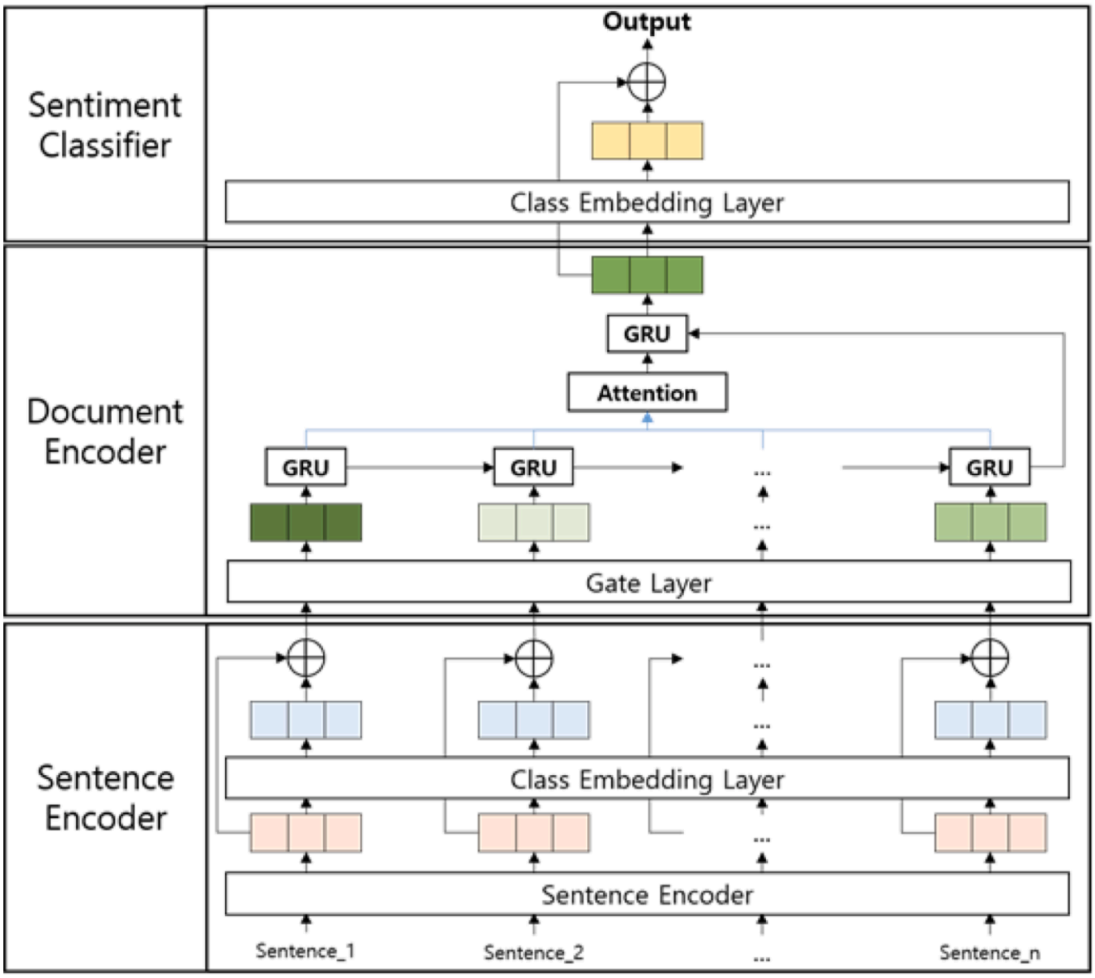

Fig. 14. Importance of Sentences Improvement of SLR (RQ6).

by the managed cycles (Dang et al., 2020; Do et al., 2019).

In the case of RecNNs, tree-structured models can better construct the syntactic performance of sentence structure. Generally, in a recursive neural model, the vector representation of each node in the tree structure is studied from the performance of all its nodes using a weight matrix W, which participates across the entire model (Do et al., 2019; Socher et al., 2013). Table 7 assesses the performance of a set of models related to this primary dataset.

The majority of the models were implemented according to DOCSA and its variants some of the applications of DOCSA, with high performance in both areas - more than 92% in the S3 with IMDB dataset and around 90% in the newspapers field as shown in Table 8. The current best model appears to be the CNN model by S3 with amazon dataset for both areas, showing that the window approach in CNN can extract relevant opinion objectives, and can overcome the long-term dependency problem.

From Tables 9, 10 and 12 it is clear that the S9 with Yelp2014, Yelp2015 and Yahoo dataset in Restaurant and Yahoo website domain gives an acceptable accuracy of around 70%.

The S14 presented a good accuracy as shown in Table 11 with Yelp2013 dataset in Restaurant and Yelp Website domain since the accuracy is around 70%.

Table 15
Improvement using Importance of Sentences.

Dataset	Domain	No.	Research	Model	Accuracy
IMDB	Amazon	1	(Choi et al., 2020)	DLSAM	Acc: 88.20%

The performance in the S32 with the Amazon data set in electronic products and the Amazon review field gave impressive results, where the accuracy reached more than 95% as shown in Table 13 while the performance was better than all previous studies in the S3 with RCV1 dataset in Reuters News domain, where the accuracy reached more than 96% as presented in Table 14.

It is also interesting that the attention mechanism can enhance the performance of RNN-based systems (eg Li et al., 2018; W. Wang et al., 2017). Fewer attempts were made to implement RecNN with lower performance, indicating that sequentially word processing may be more information

5.2. *Opportunities and Challenges*

Concerning the opportunities and challenges of SLR (RQ5) that are present in DOCSA, it is obvious that deep learning models using sentiment analysis are insufficient at this time. There are some situations where the performance of DL models is not efficient compared to other models (Do et al., 2019). An example is presented by the S33 Ghosh et al. (2016), who examined deep sentiment performance based on CNN and LSTM models. DL has a threshold over classical learning algorithms to

Table 16
Improvement using Clause and Discourse Connectives.

Dataset	Domain	No.	Research	Model	Accuracy
Malayalam	Fire Rev.	1	(Saraswathy and Lalithadevi, 2021)	CDC Gold Parse	Acc: 95.22%

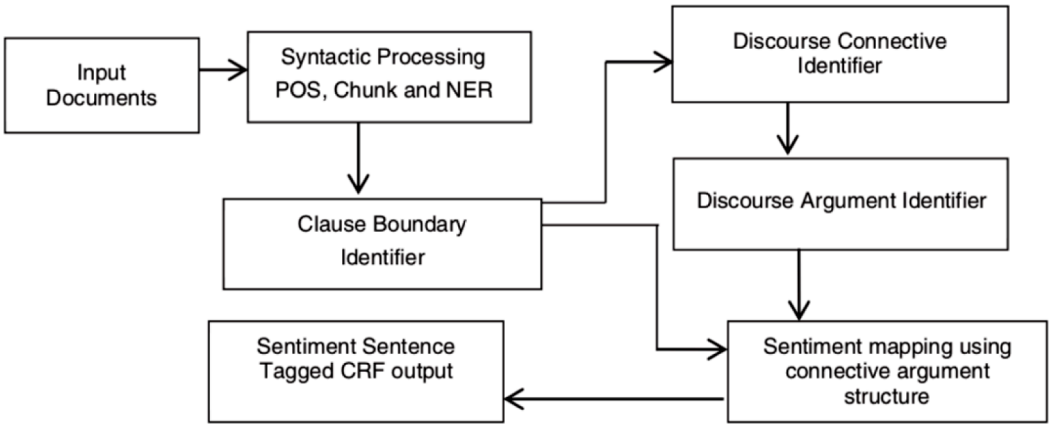

Fig. 15. Clause and Discourse Connectives Improvement of SLR (RQ6).

implement sentiment analysis because of its efficiency in treating the challenges caused by sentiment classification. Document-level sentiment classification is a basic challenge in sentiment analysis to recognize the sentiment label of a document as presented in the S14 (Wang and Liu, 2015; Wu et al., 2018).

Kim (2014) found that a model based purely on words that use max pooling achieves good performance on sentiment categorization tasks without feature engineering. Pennington et al. (2014) applied simple convolutional neural networks with static vectors and got excellent results on different datasets.

A study by Tai et al. (2015) tried to merge the concept of standard LSTM to tree- structured topologies and got superior results over a sequential LSTM model. In the S32 Liu et al. (2017), the researchers proposed three different models of DL for multi-task learning to achieve high accuracy on different datasets.

Table 17
List of All Journals

Journal	2002-2006	2007-2011	2012-2016	2017-2022
EMNLP (All)	145	596	174	3
EMNLP (Applicable)	2	1	5	3
EMNLP (Selected)	2	1	5	3
EMNLP (Application)	0	1	5	3
IEEE Intelligent Systems (All)	0	0	6	28
IEEE Intelligent Systems (Applicable)	0	0	0	2
IEEE Intelligent Systems (Selected)	0	0	0	1
IEEE Intelligent Systems (Application)	0	0	1	0
ArXiv (All)	36	100	134	61
ArXiv (Applicable)	1	1	3	5
ArXiv (Selected)	0	0	2	4
ArXiv (Application)	0	0	0	0
ACL (All)	165	351	694	960
ACL (Applicable)	2	1	7	4
ACL (Selected)	2	1	6	3
ACL (Application)	0	0	4	3
Elsevier (All)	46	100	133	144
Elsevier (Applicable)	1	2	4	4
Elsevier (Selected)	0	1	3	3
Elsevier (Application)	0	0	0	0
ACM (All)	72615	117492	134040	151537
ACM (Applicable)	2	1	3	6
ACM (Selected)	1	0	2	5
ACM (Application)	0	0	1	3
AAAI (All)	36	100	143	22
AAAI (Applicable)	0	2	7	8
AAAI (Selected)	0	0	5	3
AAAI (Application)	0	0	0	1
Springer (All)	1101	1846	3762	9951
Springer (Applicable)	0	1	0	2
Springer (Selected)	0	1	0	2
Springer (Application)	0	0	0	0
JAIR (All)	149	230	292	252
JAIR (Applicable)	1	2	1	3
JAIR (Selected)	0	1	0	2
JAIR (Application)	0	0	0	0
IEEE TKDE (All)	307	489	945	1340
IEEE TKDE (Applicable)	0	1	4	1
IEEE TKDE (Selected)	0	0	3	0
IEEE TKDE (Application)	0	0	0	0
Others (All)	5	5	20	20
Others (Applicable)	5	5	20	18
Others (Selected)	5	5	20	18
Others (Application)	0	2	6	7

Table 18
Quality Evaluation Scores of the Study

Study	Study Type	QA1	QA2	QA3	QA4	Total Score
S1	Experiment	P	P	Y	P	2.5
S2	Experiment	P	P	Y	P	2.5
S3	Experiment	Y	P	Y	Y	3.5
S4	Experiment	N	P	Y	Y	2.5
S5	Experiment	Y	P	Y	P	3
S6	Experiment	Y	Y	Y	P	3.5
S7	Experiment	P	N	Y	Y	2.5
S8	Experiment	Y	Y	Y	Y	4
S9	Experiment	Y	Y	Y	Y	4
S10	Experiment	Y	P	Y	Y	3.5
S11	Experiment	Y	Y	Y	Y	4
S12	Experiment	P	N	Y	Y	2.5
S13	Experiment	Y	Y	Y	Y	3
S14	Experiment	Y	Y	Y	Y	3
S15	Experiment	Y	Y	Y	Y	3
S16	Experiment	Y	N	Y	P	2.5
S17	Experiment	Y	P	Y	Y	3.5
S18	Experiment	Y	Y	Y	Y	4
S19	Experiment	Y	Y	Y	Y	4
S20	Experiment	N	P	Y	Y	3.5
S21	Experiment	Y	P	Y	P	3
S22	Experiment	P	Y	Y	Y	3.5
S23	Experiment	Y	Y	Y	Y	4
S24	Experiment	N	Y	Y	Y	3
S25	Experiment	Y	P	Y	Y	3.5
S26	Experiment	Y	N	Y	Y	3
S27	Experiment	Y	P	Y	P	3
S28	Experiment	Y	P	Y	Y	3.5
S29	Experiment	P	Y	Y	Y	3
S30	Experiment	N	N	Y	Y	2
S31	Experiment	Y	P	Y	Y	3
S32	Experiment	Y	P	Y	Y	3.5
S33	Experiment	Y	P	Y	Y	3.5
S34	Experiment	Y	Y	Y	Y	4
S35	Experiment	N	P	Y	Y	2.5
S36	Experiment	Y	P	Y	Y	3.5
S37	Experiment	N	P	Y	Y	2.5

Table 19
Bibliometric Analysis in Terms of Author

Study	Names	Institutions	Country	Citation	Year
S1	Rie Johnson	RJ Research Consulting	USA	467	2017
	Tong Zhang	Tencent AI Lab	China		
S2	Alexis Conneau	Facebook AI Research	USA	914	2017
	Holger Schwenk	Facebook AI Research	USA		
	Yann Le Cun	Facebook AI Research	USA		
	Loic Barrault	University of Le Mans	France		
S3	Rie Johnson	RJ Research Consulting	USA	944	2015
	Tong Zhang	Baidu Inc	China		
S4	Maryem Rhanoui	Mohammed V University	Morocco	58	2019
	Mounia Mikram	School of Information Sciences	Morocco		
	Siham Yousfi	Mohammed V University	Morocco		
	Soukaina Barzali	Mohammed V University	Morocco		
S5	Joosung Yoon	Korea University	South Korea	35	2017
	Hyeoncheol Kim	Korea University	South Korea		
S6	Duyu Tang	Harbin Institute of Technology	China	1513	2015
	Bing Qin	Harbin Institute of Technology	China		
	Ting Liu	Harbin Institute of Technology	China		
S7	Xiaodong Gu	Fudan University	China	36	2017
	Yiwei Gu	Fudan University	China		
	Haibing Wu	Fudan University	China		
S8	Duyu Tang	Harbin Institute of Technology	China	1513	2015
	Bing Qin	Harbin Institute of Technology	China		
	Ting Liu	Harbin Institute of Technology	China		
S9	Zichao Yang	Carnegie Mellon University	USA	4028	2016
	Diyi Yang	Carnegie Mellon University	USA		
	Chris Dyer	Carnegie Mellon University	USA		
	Xiaodong He	Microsoft Research	USA		
	Alex Smola	Carnegie Mellon University	USA		
	Eduard Hovy	Carnegie Mellon University	USA		
S10	Huimin Chen	Tsinghua University	China	302	2016
	Maosong Sun	Capital Normal University	China		
	Cunchao Tu	Tsinghua University	China		
	Yankai Lin	Tsinghua University	China		
	Zhiyuan Liu	Tsinghua University	China		
S11	Zi-Yi Dou	Nanjing University	China	83	2017
S12	Andrew L. Maas	Nanjing University	USA	3536	2017
	Raymond E. Daly	Nanjing University	USA		
	Peter T. Pham	Nanjing University	USA		
	Dan Huang	Nanjing University	USA		
S13	Reinald Kim Amplayo	Yonsei University	South Korea	32	2018
	Jihyeok Kim	Yonsei University	South Korea		
	Sua Sung	Yonsei University	South Korea		
	Seung-won Hwang	Yonsei University	South Korea		
S14	Zhen Wu	Nanjing University	China	79	2018
	Xin-Yu Dai	Nanjing University	China		
	Cunyan Yin	Nanjing University	China		
	Shujian Huang	Nanjing University	China		
	Jiajun Chen	Nanjing University	China		
S15	Qiao Qian	Tsinghua University	China	71	2015
	Bo Tian	Tsinghua University	China		
	Minlie Huang	Tsinghua University	China		
	Yang Liu	Samsung R&D Institute	China		
	Xuan Zhu	Samsung R&D Institute	China		
	Xiaoyan Zhu	Tsinghua University	China		
S16	Tomas Mikolov	Brno University of Technology	Czech Rep.	6216	2010
	Martin Karafiat	Brno University of Technology	Czech Rep.		
	Lukas Burget	Brno University of Technology	Czech Rep.		
	Jan Honza Cernocky	Brno University of Technology	Czech Rep.		
	Sanjeev Khudanpur	Johns Hopkins University	USA		
S17	Richard Socher	Stanford University	USA	6081	2013
	Alex Perelygin	Stanford University	USA		
	Jean Y. Wu	Stanford University	USA		
	Jason Chuang	Stanford University	USA		
	Christopher D. Manning	Stanford University	USA		
	Andrew Y. Ng	Stanford University	USA		
	Christopher Potts	Stanford University	USA		
S18	Jiacheng Xu	Fudan University	China	157	2016
	Danlu Chen	Fudan University	China		
	Xipeng Qiu	Fudan University	China		
	Xuanjing Huang	Fudan University	China		
S19	Yichun Yin	Peking University	China	60	2017
	Yangqiu Song	HKUST	Hong Kong		
	Ming Zhang	Peking University	China		
S20	Xinjie Zhou	Peking University	China	235	2016

(*continued on next page*)

Table 19 (*continued*)

Study	Names	Institutions	Country	Citation	Year
	Xiaojun Wan	Peking University	China		
	Jianguo Xiao	Peking University	China		
S21	Zheng Li	Hong Kong Univ. Sci. & Tech.	Hong Kong	160	2017
	Yu Zhang	Hong Kong Univ. Sci. & Tech.	Hong Kong		
	Ying Wei	Hong Kong Univ. Sci. & Tech.	Hong Kong		
	Yuxiang Wu	Hong Kong Univ. Sci. & Tech.	Hong Kong		
	Qiang Yang	Hong Kong Univ. Sci. & Tech.	Hong Kong		
S22	Clement Sage	Univ Lyon	France	14	2019
	Alex Aussem	Univ. Lyon	France		
	Haytham Elghazel	Univ. Lyon	France		
	Veronique Eglin	Univ. Lyon	France		
	Jeremy Espinas	Univ. Lyon	France		
S23	Jianming Zheng	National Univ. of Defense Tech.	China	15	2018
	Yupu Guo	National Univ. of Defense Tech.	China		
	Chong Feng	National Univ. of Defense Tech.	China		
	Honghui Chen	National Univ. of Defense Tech,	China		
S24	AHMAD AL-SALLAB	Cairo University	Egypt	20	2017
	RAMY BALY	American University of Beirut	Lebanon		
	HAZEM HAJJ	American University of Beirut	Lebanon		
	KHALED BASHIR SHABAN	Qatar University	Qatar		
	WASSIM EL-HAJJ	National Univ. of Defense Tech.	China		
	GILBERT BADARO	American University of Beirut	Lebanon		
S25	Richard Socher	Stanford University	USA	1549	2011
	Jeffrey Pennington	Stanford University	USA		
	Eric H. Huang	Stanford University	USA		
	Andrew Y. Ng	Stanford University	USA		
	Christopher D. Manning	Stanford University	USA		
S26	G. Preethi	SPM University	India	54	2017
	P. Venkata Krishna	SPM University	India		
	Mohammad S. Obaidat	Fordham University	USA		
	V. Saritha	Sree Vidyanikethan Eng. Col.	India		
	Sumanth Yenduri	Columbus State University	USA		
S27	Ye Yuan	Stanford University	USA	19	2015
	You Zhou	Stanford University	USA		
S28	duy-vinh Van	Ho Chi Minh City Univ. of Sci.	Vietnam	14	2017
	Thien Thai	Ho Chi Minh City Univ. of Sci.	Vietnam		
	Minh-Quoc Nghiem	Ho Chi Minh City Univ. of Sci.	Vietnam		
S29	Guangyou Zhou	Central China Normal University	China	28	2016
	Zhao Zeng	Central China Normal University	China		
	Jimmy Xiangji Huang	York University	Canada		
	Tingting He	Central China Normal University	China		
S30	Tomas Mikolov	Google Inc	USA	28072	2013
	Kai Chen	Google Inc	USA		
	Greg Corrado	Google Inc	USA		
	Jeffrey Dean	Google Inc	USA		
S31	Patrawut Ruangkanokmas	KMUTT	Thailand	52	2016
	Tiranee Achalaku	KMUTT	Thailand		
	Khajonpong Akkarajitsakul	KMUTT	Thailand		
	Jeffrey Dean	KMUTT	Thailand		
S32	Jingzhou Liu	Carnegie Mellon University	USA	429	2017
	Wei-Cheng Chang	Carnegie Mellon University	USA		
	Yuexin Wu	Carnegie Mellon University	USA		
	Yiming Yang	Carnegie Mellon University	USA		
S33	Rahul Ghosh	Indian Institute of Technology	India	31	2016
	Kumar Ravi	University of Hyderabad	India		
	Vadlamani Ravi	Center of Excellence in Analytics	India		
S34	Goebel, R	University of Alberta	Canada	N/A	2022
	Wahlster, W	German Res. Center for AI	Germany		
	Zhi-Hua Zhou	Nanjing University	China		
	Jorg Siekmann	DFKI	Germany		
S35	Kumar Ravi	University of Hyderabad	India	26	2016
	Vadlamani Ravi	Centre of Excellence in Analytics	India		
S36	Kumar Ravi	University of Hyderabad	India	94	2016
	Vadlamani Ravi	Centre of Excellence in Analytics	India		
S37	Peerapon Vateekul	Chulalongkorn University	Thailand	78	2016
	Thanabhat Koomsubha	Chulalongkorn University	Thailand		

One major challenge for DOCSA is an inability to perform well in other domains, as accuracy and performance are inadequate in sentiment analysis based on insufficient labeled data. Methods based on DL are obtaining approval. Kumar et al. (2016); Kanayama and Nasukawa (2006); Asim et al. (2019) identified that an individual DL model outperforms previous methods, reaching the highest level of performance for numerical sarcasm detection as presented in the S32.

However, DNNs were not the best for numerical sarcasm. All of this explains that there are still important challenges in terms of the application of DL methods to sentiment classification in general and for individual DOCSA tasks (Flekova et al., 2015; Riloff et al., 2013) as indicated in the S12 and S21.

Table 20
The Evaluation of the Average Quality Scores

Evaluation	2002-2006	2007-2011	2012-2016	2017-2022
Number of Researches	N/A	3	17	17
Mean Quality	N/A	3.33	55.5	53
Standard deviation	N/A	3.20	0.57	0.52

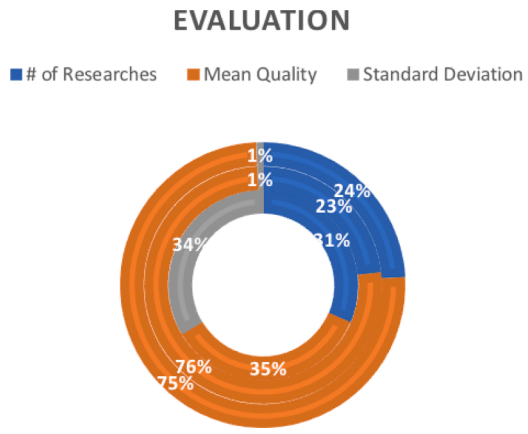

Fig. 16. Evaluation of the Quality Scores of the Resources - 1.

For DOCSA tasks, it is obvious from the discussion that while one method may implement completely in other domains, there is no assurance of having similar accuracy (Ding et al., 2017; Habimana et al., 2019) as presented in the S27 and S30.

Ghosh et al. (2016);Liu et al. (2017); Johnson and Zhang (2015); Gu et al. (2017) as presented in S33, S32, S3 and S7 outperformed regular methods for different datasets. The neural network approach is more efficient in composing the semantic illustration of text data. However, it must be noted that the SVM classifier is an extremely strong method with n-gram features compared to other baseline methods (Cheng and Tsai, 2019; Kumar et al., 2016).

5.3. Sentiment Analysis Improvement in Deep Learning at the Document Level

To classify whether a document has positive or negative polarity, each text in the document must be treated with different degrees of importance. To improve this problem and provide impressive results for documents of SLR (RQ6), many studies have been conducted, but two models are distinguished, the first using importance of sentences in the document-level sentence classification because it is based on deep neural networks, and thus the text importance degrees in documents are automatically determined by portal mechanisms. Second, it relies heavily on the clause and discourse connectives as it greatly improves the performance of sentiment analysis from the platform.

Fig. 17. Evaluation of the Quality Scores of the Resources - 2.

5.3.1. Using Importance of Sentences

In the past, researchers see sentiment analysis of the document classification task as an entry task into pre-classified emotion categories. Although there were words in the document that gave important clues to analyzing feelings, they did not, they treated the document as a group of words. In other words, they did not consider the significance of every word in the document. To positively analyze the sentiments of the document, each word in the document should be treated in order of importance. To address this issue, the researchers demonstrated a document-level word classification model that relies on deep neural networks, in which words in documents are automatically labeled with portals. To confirm their new sentiment analysis model, experiments were conducted using sentiment data sets in several different domains, such as movie reviews, hotel reviews, restaurant reviews, and music reviews. In trials, the proposed model has succeeded over earlier, more recent models that do not consider differences in the significance of words in the document. Therefore, the importance of words must be considered in classifying feelings at the document level (Choi et al., 2020).

As seen in Fig. 14 (Choi et al., 2020), the proposed model consists of three sub-modules: sentence encoder, document encoder, and sentiment classifier. The syntax encoder generates nests of each text in a given document using ALBERT (a light version of BERT). Then it enriches the texts of positive objects for sentiment chapters through class embedding layer. The document encoder relies on the significance of each text through the portal's functions and GRU and gate layer. Next, it proceeds to create an inclusion document through a weighted collection of texts based on the calculated importance. Then, the sentiment classifier enriches the document's inclusion by adding positive objects to the sentiment categories through the class embedding layer. Table 15 provides the improvement using the importance of sentences (Choi et al., 2020).

5.3.2. Using Clause and Discourse Connectives

The emotion of the text is based on the structure of the sentences and the arguments of the discourse of conjunctions. In this paper, the semantics of the beginning and end of sentences, the argument of discourse, and grammatical and semantic rules of the sentence are used to classify the feeling of the text. So the extent of the text is determined based on the semantics of the sentences, as well as the discourse that links the arguments. Since document emotion analysis based on the vocabulary of traditional sentiment analysis gives a false sense of judgment, semantic analysis on a deeper criterion is desired for correct sentiment analysis. Hence, in their work, the obvious links in Malayalam for classifying discourse arguments are considered. So a supervised method, conditional random fields, is used to determine the beginnings and endings of a sentence and discourse arguments. For the study, more than 1,000 emotional sentences from Malayalam documents were analyzed. So the experimental results show that the structure of the sentences and discourse arguments improve the performance of sentiment analysis of the platform. Table 16 provides the improvement using the importance of sentences (Saraswathy and Lalithadevi, 2021).

The proposed work as shown in Fig. 15 (Saraswathy and Lalithadevi, 2021) is built based on the beginnings and endings of sentences so the work consists of five steps: 1) determining the sentence boundaries of the sentence, 2) determining the type of sentence POS, chunk and NER, 3) defining the discourse links identifier, 4) determining the arguments of the conjunctions, 5) defining the sign of feelings using connective argument structure. Here, text boundaries allocate the grammatical rules, while the discourse argument allocates the semantic information of the sentence. 6) Then these two characteristics are combined to analyze the feeling of the sentence using tagged CRF output. The speech mark's argument can be a phrase or a sentence. The discourse

relationship can also be classified as internal or implicit. In sentence-within-sentence relationships, the feeling of the judgment depends on the feeling of the linked sentence.

6. Evaluation

We classified 37 relevant studies into the sources we searched for their application, as shown in Table 17. So this section summarizes the results of the study. The quality evaluation scores for studies between 2002 and 2022 are shown in Table 18 (Kitchenhamy et al., 2009). The bibliometric analysis is also presented in Table 19. The average quality scores for studies between 2002 and 2022 are shown in Table 20 (Kitchenhamy et al., 2009).

Table 19 shows the bibliometric analysis in terms of author names, author institutions, author countries, citations and year of the publications.

Table 20 indicates that the number of studies published in the cited years was quite stable (Kitchenhamy et al., 2009). We found that 17 studies between 2017 and 2022 were highly rated. Thus, the average quality result appears to be steadily rising. Furthermore, the standard deviation from the years 2017 to 2022 was quite high about 0.52 and shows that the source data are clustered closely around the mean (more reliable) as shown in Fig. 16 and Fig. 17. Since the highest number of researchers is 31% between 2017 and 2022.

Regarding where SLRs are published, EMNLP, ACL, Elsevier, ACM and others published 3 studies or more, AAAI published 1 study. Thus, IEEE Intelligent Systems, ArXiv and IEEE TKDE attempt to stimulate the deployment of SLRs.

We noted that Springer, JAIR and IEEE TKDE libraries did not include any relevant DOCSA studies, even though these journals published a systematic review of the literature on the topic of deep learning techniques. An automated search of these surveys using these libraries from 2002 to 2022 found no application-related surveys.

Regarding the topic of the articles, seven of them were related to CNN applications and 16 studies were related to RNN. As for the RecNN, the number of studies was 5. With regards to DBN, it was 3 studies. Finally, the number of studies on HNN was 6.

7. Conclusion

After document content emerged as a spacious source of particular information, there have been powerful efforts to classify, recognize, and analyze the sentiment polarity of documents using different languages. This analysis offers an inclusive SLR of significant DL methods and provides an accurate contrast of these methods for sentiment analysis at the document level. For this analysis, more than 30 studies were outlined and categorized according to their essential structure and classification functions. General methods contain the standard and alternative methods of CNN, LSTM, RNN, RecNN, and HNN. For the performance of methods, research has covered pre-trained documents at the document level in the form of language processing techniques as well as examining idea-based information.

However, from the SLR of the most recent findings in document-level sentiment analysis and introduced in this research, it is obvious that DOCSA and deep learning are still in the first steps. Given the relationship between document and opinion, recovered performance can be gained by common extraction and classification of document, class, and sentiment. On the other hand, many powerful studies opted to implement only document extraction, and those who simultaneously implement document detection and sentiment analysis, have not yet accomplished optimal accuracy. Subsequently, there is a need for a joint method that can assume both tasks and make more extensive sentiment analysis at the document level.

The robust neural network model for document-level sentiment analysis determines the importance scores of texts in documents using portal functions taken from aggregate training data.

Therefore, this SLR first presents the background of sentiment analysis, including its applications, levels, and tasks of sentiment analysis. Likewise, it introduces traditional methods of sentiment analysis and their drawbacks. We mainly discuss deep learning approaches and their applications in various tasks of full document-level, multi-modal sentiment analysis. Moreover, we take up the exploration of those applications previously and currently presented, and present a performance analysis of the results they achieved in real-world datasets for each task. Finally, we highlight current issues that need to be addressed and make suggestions for improvement and performance including the use of new paradigms such as the use of importance of sentences taken from aggregate training data as well as the use of Clause and Discourse Connectives.

Furthermore, the standard deviation from the years 2017 to 2021 was quite high about 0.52 and shows that the source data are clustered closely around the mean (more reliable). Therefore, in recent years there are many research and experiments as well as publication in the DOCSA compared to previous years.

Future studies would improve from a more concept-focused method to join awareness standards with deep learning models. Furthermore, we will highlight current issues that have been presented in this research that should be addressed and provide suggestions for improvement including the use of new cues, new paradigms, and the use of word embedding models for sentiments in the document levels.

More efforts will be made to highlight the best performance metrics on sentiment analysis in the document levels and to benefit from the positive results of the Aspect levels. Moreover, the most processed and researched was done in the English language. At present, there are very few studies done on sentiment classification for other languages such as Arabic. We will expand to present more comprehensive studies of document-based sentiment analysis of Arabic texts.

We will also focus on exploring hybrid approaches of DL, to integrate different models and techniques to extend the study to both new approaches and to enhance the accuracy of sentiment classification achieved by individual models. Therefore, the results of hybrid models' reliability and processing time using several types of data will be shown.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. List of Abbreviations

CNN: Convolutional Neural Network.
DOCSA: Document-Based Sentiment Analysis.
RBM+PNN: Restricted Boltzmann Machine + Probabilistic Neural Network.
DBLSTM: Deep Bi-directional Long Short- Term Memory Neural Networks.
BPN: Back Propagation Neural Network.
DCNN: Deep Convolutional Neural Network.
LSTM: Long Short-Term Memory.
RecNN: Recursive Neural Networks.
RNN: Recurrent Neural Networks.
NLP: Natural Language Processing.
ANN: Artificial Neural Network.
DNN: Deep Neural Networks.
DBNs: Deep belief networks.
GRNN: Generalized regression neural network.
SSTb: Stanford Sentiment Tree-bank.
IMDB: Internet Movie Database.
UPA: User Product Attention.
UPNN: User Product Neural Network.
LSA: Latent Semantic Analysis.

SVM: Support Vector Machine.
NB: Naïve Bayes.
dJST: dynamic Joint Sentiment-Topic Model.
HNN: Hybrid Neural Networks.
CBOW: Continuous Bag of Words.
Glove: Global Vectors for Word Representation.
POS: Part-Of-Speech.
NP: Noun Phrase.
VP: Verb Phrase.
ADJP: Adjective Phrase.
ADVP: Adverb Phrase.
CRF: Conditional Random Fields.
DPCNN: Deep Pyramid Convolutional Neural Network.
VDCNN: Very Deep Convolutional Neural Network.
GRU: Gated Recurrent Units.
RRNN: Regular Recurrent Neural Networks.
BRNN: Bidirectional Recurrent Neural Networks.
HAN: Hierarchical Awareness Network.
CBA: Cognition-Based Attention.
LA: Local Attention.
HCSC: Hybrid Contextualized Sentiment Classifier.
HUAPA: Hierarchal User Attention Product.
Acc: Accuracy.
Per: Performance.
TG: Transformation Gated.
CLSTM : Contextual LSTM.
AROMA: A Recursive Deep Learning Model for Opinion Mining in Arabic.
RAE: Recursive Autoencoders.
SemEval: International Workshop on Semantic Evaluation.
DL: Deep Learning.
NSC: Neural Sentiment Classification.
XMTC: Extreme Multi-label Text Classification.
WSDNN: Weakly Shared Deep Neural Network
DBNWV: Deep Belief Network Word Vector
DBNFS: Deep Belief Network Feature Selection
DLSAM: Document-Level Sentiment Analysis Model
CDC: Clause and Discourse Connectives
SLR: Systematic Literature Review

References

Ain, Q., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: A review. *International Journal of Advanced Computer Science and Applications, 8*, 424–433.

Al-Sallab, R., Baly, A., Hajj, H., Shaban, K., El-Hajj, W., & Badaro, G. (2017). Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Proces, 16*, 1–20.https://doi.org/10.1145/3086575

Amplayo, R., Kim, J., Sung, S., & Hwang, S. (2018). Cold-start aware user and product attention for sentiment classification. *Proc. of the 56th annual meeting of the association for computational linguistics, acl* (pp. 2535–2544).

Asim, M., Khan, M., Malik, M., Dengel, A., & Ahmed, S. (2019). A robust hybrid approach for textual document classification. *arXiv:1909.05478v1, CS.CL*.

Behdenna, S., Barigou, F., & Belalem, G. (2018). Document level sentiment analysis: A survey. *EAI Endorsed Transactions on Context-aware Systems and Applications, 4*.https://doi.org/10.4108/eai.14-3-2018.154339

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.

Bengio, Y., Schwenk, H., Senecal, J., Morin, F., & Gauvain, J. (2006). Neural probabilistic language models. *Studies in Fuzziness and Soft Computing, 194*, 137–186.

Bongirwar, V. (2015). A survey on sentence level sentiment analysis. *International Journal of Computer Science Trends and Technology (IJCST), 3*, 110–113.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 32*, 74–80.https://doi.org/10.1109/MIS.2017.4531228

Chaturvedi, N. (2022). Recurrent neural network with keras. https://medium.datadriveninvestor.com/recurrent-neural-network-with-keras-b5b5f6fe5187, Accessed: 2022.03.05.

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. *Expert Systems with Applications, Elesvier, 27*, 221–230.https://doi.org/10.1016/j.eswa.2016.10.065

Chen, T., Xu, R., He, Y., Xia, Y., & Wang, X. (2016). Learning user and product distributed representations using a sequence model for sentiment analysis. *article*.

Cheng, L., & Tsai, S. (2019). Deep learning for automated sentiment analysis of social media. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1001–1004.https://doi.org/10.1145/3341161.3344821

Choi, G., Oh, S., & Kim, H. (2020). Improving document-level sentiment classification using importance of sentences. *Entropy (Basel, Switzerland), 22*, 1–12.https://doi.org/10.3390/e22121336

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS 2014 Deep Learning and Representation Learning Workshop, arXiv:1412.3555, CS.CL*, 1–9.

Conneau, A., Schwenk, H., & L, B. (2017). Very deep convolutional networks for text classification. *Proc. of the 15th conference of the european chapter of the association for computational linguistics*.11107–1116

Costa, M. (2018). From feature to paradigm: Deep learning in machine translation. *Journal of Artificial Intelligence Research (Jair), 61*, 947–974.https://doi.org/10.1613/jair.1.11198

Dang, N., Moreno-Garcia, M., & Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics - MDPI, 9*, 195–225.https://doi.org/10.3390/electronics9030483/

Dhande, L., & Patnaik, G. (2014). Analyzing sentiment of movie review data using naive bayes neural classifier. *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS), 3*, 313–320.ISSN 2278-6856

Ding, Y., Yu, J., & Jing, J. (2017). Recurrent neural networks with auxiliary labels for cross-domain opinion target extraction. *Proc. of the 31st aaai conference on artificial intelligence* (pp. 3436–3442).

Do, H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems With Applications, available on ELSEVIER, 118*, 272–299.https://doi.org/10.1016/j.eswa.2018.10.003

Dou, Z. (2017). Capturing user and product information for document level sentiment analysis with deep memory network. *Proc. of the 2017 conference on empirical methods in natural language processing (emnlp)* (pp. 521–526).https://doi.org/10.18653/v1/D17-1054

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 17*, 195–225.https://doi.org/10.1023/A:1022699029236

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. *Proc. of the 14th acm international conference on information and knowledge managemen (cikm '05)* (pp. 617–624).https://doi.org/10.1145/1099554.1099713

Flekova, L., Preotiuc-Pietro, D., & Ruppert, E. (2015). Analysing domain suitability of a sentiment lexicon by identifying distributionally bipolar words. *Proc. of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis, acl* (pp. 77–84).https://doi.org/10.1515/popets-2015-0023

Geron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc.

Ghosh, R., Ravi, K., & Ravi, V. (2016). A novel deep learning architecture for sentiment classification. *3rd IEEE Int. Conf. Recent Adv. Inf. Technology*, 511–516.

Ghosh, S. (2022). Invoice information extraction using ocr and deep learning. https://medium.com/analytics-vidhya/invoice-information-extraction-using-ocr-and-deep-learning-b79464f54d69, Accessed: 2022.03.01.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proc. of the 28th international conference on machine learning*.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-scale sentiment analysis for news and blogs (system demonstration). *Proc. of the international conference on weblogs and social media (icwsm'07)*.

Goebel, R., Wahlster, W., Zhou, Z., & Siekmann, J. (2022). *Integrated Uncertainty in Knowledge Modelling and Decision Making, Lecture Notes in Artificial Intelligence*. Springer.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research (Jair), 57*, 345–420.

Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publisher.

Graves, A. (2008). *Supervised sequence labelling with recurrent neural networks* (vol. 12). Studies in Computational Intelligence, Springer.

Gu, X., Gu, Y., & Wu, H. (2017). Cascaded convolutional neural networks for aspect-based opinion summary. *Neural Processing Letters, 46*, 581–594.https://doi.org/10.1007/s11063-017-9605-7

Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2019). Sentiment analysis using deep learning approaches: an overview. *Science China Information Sciences, Springer, 63*.

He, Y., Lin, C., Gao, W., & Wong, K. (2012). Tracking sentiment and topic dynamics from social media. *Proc. of the sixth international aaai conference on weblogs and social media* (pp. 1483–1486).

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proc. of the 2004 acm sigkdd international conference on knowledge discovery and data mining - kdd '04* (pp. 168–177).https://doi.org/10.1145/1014052.1014073

JChen, H., Sun, M., Tu, C., Lin, Y., & Liu, Z. (2016). Neural sentiment classification with user and product attention. *Proc. of the 2016 conference on empirical methods in natural language processing (acl)* (pp. 1650–1659).

Jebbara, S., & Cimiano, P. (2017). Aspect-based relational sentiment analysis using a stacked neural network architecture. *Computation and Language (cs.CL), arXiv: 1709.06309*.https://doi.org/10.3233/978-1-61499-672-9-1123

Johnson, R., & Zhang, T. (2015). Effective use of word order for text categorization with convolutional neural networks. *Proc. of the 2015 conference of the north american*

chapter of the association for computational linguistics: Human language technologies (acl) (pp. 103–112).https://doi.org/10.3115/v1/N15-1011
Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. *Proc. of the 55th annual meeting of the association for computational linguistics (acl)*.11562–1570
Kanayama, H., & Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. *Proc. of the 2006 conference on emnlp* (pp. 355–363). https://doi.org/10.1515/popets-2015-0023
Karani, D. (2018). Introduction to word embedding and word2vec. *towards data science*. https://towardsdatascience.com
Kim, Y. (2014). Convolutional neural networks for sentence classification. *Arxiv, CS.CL*, 23–31.
Kitchenhamy, B., Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - a systematic literature review. *ELSEVIER, 51*, 7–15.https://doi.org/110.1016/j.infsof.2008.09.009
Kolkur, S., Dantal, G., & Mahe, R. (2015). Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161, 5*, 768–770.
Kulshrestha, R. (2019). Nlp 101: Word2vec - skip-gram and cbow. *towards data science*. https://towardsdatascience.com
Kumar, A., Kohail, S., Kumar, A., Ekbal, A., & Biemann, C. (2016). Beyond sentiment lexicon: combining domain dependency and distributional semantics features for aspect based sentiment analysis. *Proc. of the 10th international workshop on semantic evaluation* (pp. 1129–1135).https://doi.org/10.1515/popets-2015-0023
Lai, S., Xu, L., Liu, K., Li, M., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. *Proc. of the twenty-ninth aaai conference on artificial intelligence, aaai'15* (pp. 2267–2273).
Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents, *vol. 32. Proc. of the 31st international conference on machine learning*.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.
Li, Z., Zhang, Y., Wei, Y., Wu, Y., & Yang, Q. (2017). End-to-end adversarial memory network for cross-domain sentiment classification. *Proc. of the twenty-sixth international joint conference on artificial intelligence (ijcai-17)* (pp. 2237–2243).
Liu, J., Chang, W., Wu, Y., & Yang, Y. (2017). Deep learning for extreme multi-label text classification. *Proc. of of the 40th international acm sigir conference on research and development in information retrieval* (pp. 115–124).https://doi.org/10.1145/3155133.3155158
Long, Y., Lu, Q., Xiang, R., Li, M., & Huang, C. (2017). Learning word vectors for sentiment analysis. *Proc. of the 2017 conference on empirical methods in natural language processing, emnlp* (pp. 462–471).https://dx.doi.org/10.18653/v1/D17-1048
Ma, Y., Peng, H., & Cambria, E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.https://sentic.net/sentic-lstm.pdf
Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proc. of the 49th annual meeting of the association for computational linguistics: Human language technologies, acl* (pp. 142–150).
Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proc. of the international conference on learning representations - iclr 2013* (pp. 1–12).https://doi.org/10.1162/153244303322533223
Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *INTERSPEECH, ISCA*, 1045–1048.
Moraes, R., Valiati, J., & Neto, W. (2013). Document-level sentiment classification: An empirical comparison between svm and ann. *DBLP - Elsevier, Expert Systems with Applications, 40*, 621–633.https://doi.org/10.1016/j.eswa.2012.07.059
Morency, L., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. *Proc. of international conference on multimodal interfaces* (pp. 169–176).https://doi.org/10.1145/2070481.2070509
Nguyen, H., & Le Nguyen, M. (2017). A deep neural architecture for sentence-level sentiment classification in twitter social networking. *arXiv:1706.08032v1 [cs.CL]*.
Pang, B., & Lee, L. (2002). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. *Proc. of the 43rd annual meeting of the acl* (pp. 115–124).
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proc. of conference on empirical methods in natural language processing (emnlp)*.
Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation, *vol. 1. Proc. of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).https://doi.org/10.3115/v1/D14-1162
Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, Elesvier, 108*, 42–49. https://doi.org/10.1016/j.knosys.2016.06.009
Preethi, G., Krishna, P., Obaidat, M., Saritha, V., & Yenduri, S. (2017). Application of deep learning to sentiment analysis for recommender system on cloud. *Proc. of the 2017 international conference on computer, information and telecommunication systems (cits)* (pp. 93–97).
Qian, Q., Tian, B., Huang, M., Liu, Y., Zhu, X., & Zhu, X. (2015). Learning tag embeddings and tag-specific composition functions in recursive neural network. *Proc. of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, acl, ijcnlp* (pp. 1365–1374).
Rana, T., & Cheah, Y. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review, 46*, 459–483.
Raoa, G., Huanga, W., Fengb, Z., & Cong, Q. (2018). Lstm with sentence representations for document-level sentiment classification. *ScienceDirect - Elsevier, Neurocomputing, 308*, 49–57.https://doi.org/10.1016/j.neucom.2018.04.045
Ravi, K., & Ravi, V. (2016). Sentiment classification of hinglish text. *2016 3rd Int. Conf. Recent Adv. Inf. Technol. RAIT 2016*, 641–645.
Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. (2019). A cnn-bilstm model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction - MDPI, 1*, 832–847.https://doi.org/10.3390/make1030048
Riloff, E., Qadir, A., Surve, P., Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *Proc. of the 2013 conference on empirical methods in natural language processing (emnlp)* (pp. 704–714).
Rong, W., Nie, Y., Ouyang, Y., Peng, B., & Xiong, Z. (2014). Auto-encoder based bagging architecture for sentiment anal. *Journal of Visual Languages and Computing, 25*, 840–849.https://doi.org/10.1016/j.jvlc.2014.09.005
Ruangkanokmas, P., Achalakul, T., & Akkarajitsakul, K. (2016). Deep belief networks with feature selection for sentiment classification. *Proc. of the 7th international conference on intelligent systems, modelling and simulation - ieee* (pp. 9–14).https://doi.org/10.1109/ISMS.2016.9
Sage, C., Aussem, A., Elghazel, H., Eglin, V., & Espinas, J. (2019). Recurrent neural network approach for table field extraction in business documents. *International Conference on Document Analysis and Recognition, ICDAR*.
Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *Proc. of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 69–78).
Saraswathy, K., & Lalithadevi, S. (2021). Enhancement of sentiment analysis using clause and discourse connectives. *Computers, Materials and Continua, 68*, 1983–1999. https://doi.org/10.32604/cmc.2021.015661
Schouten, K., & Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering, 28*, 813–830.https://doi.org/10.1109/TKDE.2015.2485209
Schuster, M., & Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE TRANSACTIONS ON SIGNAL PROCESSING, 45*.
Sharma, A., & Dey, S. (2012). A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. *ACM SIGAPP Applied Computing Review, 12*, 67–75.https://doi.org/10.1145/2432546.2432552
Sharma, R., Nigam, S., & Jain, R. (2014). Opinion mining of movie reviews at document level. *International Journal on Information Theory (IJIT), 3*, 13–21.https://doi.org/10.5121/ijit.2014.3302
Socher, R., Pennington, J., Huang, E., Ng, A., & Manning, C. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proc. of the 2011 conference on empirical methods in natural language processing (emnlp)* (pp. 151–161). https://doi.org/10.18653/v1/D17-1217
Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proc. of the 2013 conference on empirical methods in natural language processing, emnlp* (pp. 1631–1642).
Stone, X., Bales, F., & Namenwirth, Z. (2007). The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Systems Research and Behavioral Science, 7*, 484–498.
Tai, K. S., Socher, R., & Manning, C. (2015). Improved semantic representations from tree-structured long short-term memory networks. *Proc. of acl, conf.* (pp. 1556–1566).https://doi.org/10.1515/popets-2015-0023
Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *DBLP - Elsevier, Expert Systems with Applications, 34*, 2622–2629.https://doi.org/10.1016/j.eswa.2007.05.028
Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. *Proc. of the 2015 conference on empirical methods in natural language processing* (pp. 1422–1432).
Tang, D., Wei, F., Qin, B., Yang, N., Liu, T., & Zhou, M. (2016). Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering, 28*, 496–509.https://doi.org/10.1109/TKDE.2015.2489653
TensorFlow (2022). word2vec tutorial. https://www.tensorflow.org/tutorials/text/word2vec, Accessed: 2022-04-01.
Timmaraju, A., & Khanna, V. (2015). Sentiment analysis on movie reviews using recursive and recurrent neural network architectures. *Proc. of corpus id: 17622339* (pp. 1–6).
Toyer, S., Thiebaux, S., Trevizan, F., & Xie, L. (2020). Asnets: Deep learning for generalised planning. *Journal of Artificial Intelligence Research (Jair), 68*, 1–68. https://doi.org/10.1613/jair.1.11633
Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proc. of the 40th annual meeting of the association for computational linguistics (acl)* (pp. 417–424).
Van, V., Thai, T., & Nghiem, M. (2017). Combining convolution and recursive neural networks for sentiment analysis. *Proc. of eighth international symposium on information and communi- cation technology, acm*.https://doi.org/10.1145/3155133.3155158
Vateekul, P., & Koomsubha, T. (2016). A study of sentiment analysis using deep learning techniques on thai twitter data. *Conference: 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1–6.https://doi.org/10.1109/JCSSE.2016.7748849
Wang, B., & Liu, M. (2015). Deep learning for aspect- based sentiment analysis. https://cs224d.stanford.edu/reports/WangBo.pdf/.
Wikipedia (2022). Recursive neural network. https://en.wikipedia.org/wiki/Recursive_neural_network, Accessed: 2022.03.15.
Wu, Z., Dai, X., Yin, C., Huang, S., & Chen, J. (2018). Improving review representations with user attention and product attention for sentiment classification. *Proc. of the thirty-second aaai conference on artificial intelligence (aaai-18)* (pp. 5989–5996).
Xia, R., Xu, F., Zong, C., Li, Q., Qi, Y., & Li, T. (2015). Dual sentiment analysis: Considering two sides of one review. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 27*.https://doi.org/10.1109/TKDE.2015.2407371

Xu, J., Chen, D., Qiu, X., & Huang, X. (2016). Cached long short-term memory neural networks for document-level sentiment classification. *arXiv:1610.04989 [cs.CL], EMNLP2016*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proc. of the conference of the north american chapter of the association for computational linguistics: Human language technologies, (acl)* (pp. 1480–1489).https://dx.doi.org/10.18653/v1/N16-1174

Yin, Y., Song, Y., & Zhang, M. (2017). Document-level multi-aspect sentiment classification as machine comprehension. *Proc. of the 2017 conference on empirical methods in natural language processing (emnlp)* (pp. 2044–2054).https://doi.org/10.18653/v1/D17-1217

Yoon, J., & Kim, H. (2017). Multi-channel lexicon integrated cnn-bilstm models for sentiment analysis. *Proc. of the 29th conference on computational linguistics and speech processing (rocling 2017)* (pp. 244–253).

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine, 13*, 55–75.https://doi.org/10.1109/MCI.2018.2840738

Yuan, Y., & Zhou, Y. (2015). Twitter sentiment analysis with recursive neural networks. *Corpus ID: 15847496, 16*, 1–8.

Zhai, S., & Zhang, Z. (2016). Semisupervised autoencoder for sentiment analysis. *Proc. of the thirtieth aaai conference on artificial intelligence* (pp. 1394–1400).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *ArXiv, Wires Data Mining and Knowledge Discovery, CS.CL, 8*.https://doi.org/10.1002/widm.1253

Zhao, Y., Dong, S., & Li, L. (2014). Sentiment analysis on news comments based on supervised learning method. *International Journal of Multimedia and Ubiquitous Engineering, 9*, 333–346.https://doi.org/10.14257/ijmue.2014.9.7.28

Zhao, Y., Qin, B., & Liu, T. (2017). Encoding syntactic representations with a neural network for sentiment collocation extraction. *Science China Information Sciences, available on Springer, 60*.https://doi.org/10.1007/s11432-016-9229-y

Zharmagambetov, A., & Pak, A. (2015). Sentiment analysis of a document using deep learning approach and decision trees. *Electronics Computer and Computation (ICECCO)*.

Zheng, J., Guo, Y., Feng, C., & Chen, H. (2018). A hierarchical neural-network-based document representation approach for text classification. *Mathematical Problems in Engineering, 2018*.https://doi.org/10.1155/2018/7987691

Zhou, G., Zeng, Z., Huang, J., & He, T. (2016a). Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks. *Proc. of the sigir 16 - acm* (pp. 245–254).https://doi.org/10.1145/2911451.2911490

Zhou, X., Wan, X., & Xiao, J. (2016b). Attention-based lstm network for cross-lingual sentiment classification. *Proc. of the 2016 conference on empirical methods in natural language processing (emnlp)* (pp. 247–256).https://dx.doi.org/10.18653/v1/D16-1024

Faisal Alshuwaier is an academic researcher in National Center for Data Analytic and Artificial Intelligence at KACST.

Ali Areshey is a programs developer in Communication and Information Technology Research Institute at KACST.

Dr. Josiah Poon is a senior lecturer in school of computer science at the University of Sydney.