Multimodal Sentiment Analysis: A Survey of Methods, Trends and Challenges

RINGKI DAS and THOUDAM DOREN SINGH, National Institute of Technology, Silchar, India

Sentiment analysis has come long way since it was introduced as a natural language processing task nearly 20 years ago. Sentiment analysis aims to extract the underlying attitudes and opinions toward an entity. It has become a powerful tool used by governments, businesses, medicine, marketing etc. The traditional sentiment analysis model focuses mainly on text content. However, technological advances have allowed people to express their opinions and feelings through audio, image and video channels. As a result, sentiment analysis is shifting from unimodality to multimodality. Multimodal sentiment analysis brings new opportunities with the rapid increase of sentiment analysis as complementary data streams enable improved and deeper sentiment detection which goes beyond text-based analysis. Audio and video channels are included in multimodal sentiment analysis in terms of broadness. People have been working on different approaches to improve sentiment analysis system performance by employing complex deep neural architectures. Recently, sentiment analysis has achieved significant success using the transformer-based model. This paper presents a comprehensive study of different sentiment analysis approaches, applications, challenges and resources then concludes that it holds tremendous potential. The primary motivation of this survey is to highlight changing trends in the unimodality to multimodality for solving sentiment analysis tasks.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Multimodal sentiment analysis**;

Additional Key Words and Phrases: Multimodal sentiment analysis, text sentiment analysis, image sentiment analysis, audio sentiment analysis, transfer learning

1 INTRODUCTION

Sentiment analysis, also known as opinion mining, aims to evoke opinions or sentiments of a person who encounters a specific topic, person, or entity [1]. Sentiment analysis can be divided into two tasks: opinion mining and emotion mining [2]. The term sentiment analysis perhaps first appeared in Nasukawa et al. [3] and the term opinion mining first appeared in Dave et al. [4]. Sentiment analysis and opinion mining mainly focus on opinions that express or imply positive or negative sentiments. For example, "Janny likes the camera of the phone," from the technical definition of opinion mining, Janny is the *opinion holder*, the camera is the *entity*, and her sentiment towards the *aspect* is positive. From the early stage of sentiment analysis, it has become an influential field for researchers, and, with time, it has established widespread applications in real life [5], [6]. Various applications of sentiment analysis found in business, government, biomedical and recommendation systems have become well-acknowledged by researchers, companies, governments, healthcare, education and other organizations. Because of covering most of the topics in natural language processing (NLP), Liu [7] describes sentiment analysis as "mini-NLP." In the same way, Cambria et al. [8] present sentiment analysis as a big suitcase of problem-solving tasks like semantic, open syntactic and pragmatic.

Authors' address: Ringki Das, ringkidas@gmail.com; Thoudam Doren Singh, thoudam.doren@gmail.com, National Institute of Technology, Silchar, India.

0360-0300/2023/3-ART $15.00
https://doi.org/10.1145/3586075

A few years ago, people used to express their thought only using text data. Only utilizing the text modality sometimes becomes a barrier to proper sentiment prediction. Nowadays, sentiment analysis approaches have started to incorporate the information from the text and other modalities like audio and visual data [9]. Sentiment prediction from other modalities, such as speech and visual, are considered robust platforms for their fantastic performance [10]. Therefore, multimodal sentiment analysis integrates different modalities and ignores a single text or image sentiment analysis model [11]. Recent studies have attempted to recognize sentiment expressed in multimedia through multimodal signals, such as visual, audio, and textual information. Thus, the internet has gradually moved from a text community to a multimedia community. It brings revolutionary changes to the sentiment analysis domain to handle various applications. Even though multimodal sentiment analysis is still in its infancy, it will take industry investment and more research to demonstrate its full potential. Many research directions are open to being widely studied, like the cause of sentiment, sentiment reasoning, understanding the motive, sentiment dialogue generation, etc.

The major trends in sentiment analysis are lexical-based, machine learning-based, and deep learning-based approaches [12]. Deep learning approaches have led to breakthroughs in sentiment analysis tasks. But the deep learning approaches are hungry for a massive amount of data and are also context-dependent. The lexical, contextual, and syntactic features have been widely embraced in state-of-art works. Because of the emergence of contextualized networks and embeddings such as BERT, we can efficiently compute a better representation of the extracted features. Equipped with thousands of parameters, transformer-based networks such as BERT [13], RoBERT [14], and their variants have pushed the existing technology to new heights. Training the data with modern architecture exploits sentiment analysis research in a new direction, such as multimodal sentiment learning, transfer learning, multilingual sentiment analysis, multidomain sentiment classification, etc.

Ortis et al. [9] reviewed pertinent publications and presented an exhaustive overview of the visual sentiment analysis field. They discussed the main issues, pros, cons of each approach, dataset and techniques related to visual sentiment analysis. The author described the design principles of visual sentiment analysis systems from emotional models, dataset definition and feature design points of view. A formalization of the problem is discussed by considering different granularity levels and the components that can affect the sentiment toward an image. Challenges, evaluation parameters and applications of visual sentiment analysis are described in the review study. Zhao et al. [10] reported the existing methods for image sentiment analysis including two main challenges affective gap and perception subjectivity. They introduced the key emotion representation models which are widely employed in affective image content analysis (AICA). Available datasets for performing evaluation and emotion feature extraction methods are also briefly described. Ortis et al. [11] introduced the research field of image sentiment analysis, reviewed the related problems, provided an in-depth overview of current research progress also discussed the major issues, dataset and outlined the new opportunities and challenges in this area. A generalizable analysis of the problem is presented by identifying and analyzing the components that affect the sentiment toward an image. Soleymani et al. [1] performed an extensive study on multimodal sentiment analysis (MSA) and reviewed recent developments, including spoken reviews, images, video blogs, human-machine and human-human interactions. The challenges and opportunities of this emerging field are also discussed. They presented an overview of the concept and goals of multimodal sentiment analysis. The review work demonstrated that multimodal sentiment analysis is becoming an important research area in the natural language processing domain. Li et al. [15] presented an overview of social media topics and described sentiment analysis and opinion-mining algorithms for social multimedia. They conducted a brief review of textual sentiment analysis for social media and a comprehensive survey of visual and multimodal sentiment analysis. They summarized the existing benchmark datasets and discussed the future research trends and potential directions for multimedia sentiment analysis. Several issues and challenges on social media analytics are reported by Singh et al. [16]. Huddar et al. [17] also stated the approaches, problems and challenges in multimodal sentiment analysis. A detailed survey on multimodal sentiment analysis consisting of feature extraction algorithms, data fusion methods and classification

techniques was presented by Chandrasekaran et al. [18]. Recently, Kaur et al. [19] surveyed multimodal sentiment analysis including opportunities and limitations of MSA. Gandhi et al. [20] reported a survey paper to examine the primary taxonomy and newly released multimodal fusion architectures.

The primary goal of this paper is to focus on the journey of sentiment analysis from unimodality to multimodality. To begin with, some of the commonly used sentiment analysis techniques are categorized and outlined in brief detail followed by a comprehensive overview of the latest developments in the field of multimodal sentiment analysis. Apart from multiple modalities, unimodal features including text, image and audio information with various approaches are also discussed. Multimodality has been explored for other NLP tasks such as machine translation [21]. The paper is reviewed in several broad dimensions, viz. sentiment classification, subjectivity classification, lexicon creation, extraction of sentiment-related information from the visual, audio and multimodal content. Further, the article also covers the categories according to the contributions to different sentiment analysis approaches to improve the system's performance by employing complex deep neural and transformer-based architectures. One of the most important contributions of the paper is to present a list of available sentiment lexicons and public datasets for sentiment analysis. The challenges, applications and evaluating measures are also included to monitor the new trending research. This survey aims to guide a neophyte researcher to address new challenges and perceive the most common challenges to look forward to a new solution of multimodal sentiment analysis. Modalities of sentiment analysis are shown in Fig. 1.

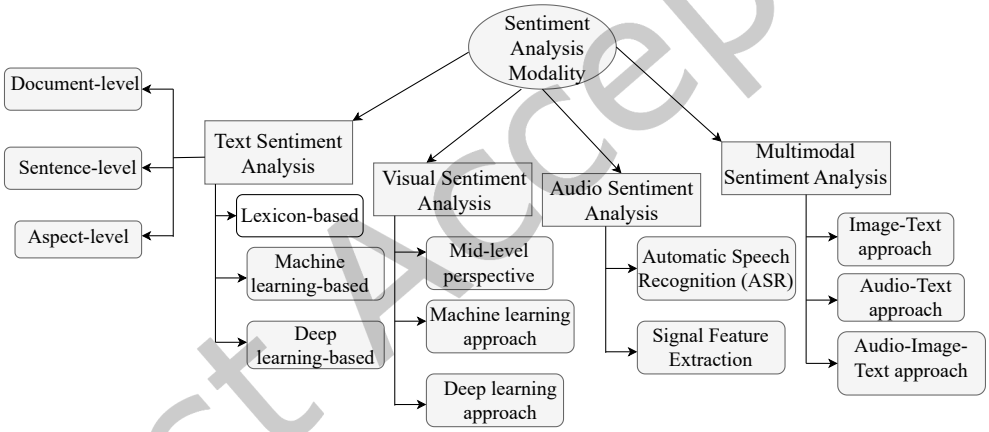

Fig. 1. Modalities of sentiment analysis

The significant contributions of this survey are outlined as follows:

- We discuss the journey from unimodal sentiment analysis to multimodal sentiment analysis research, including image, text, audio, and video information.
- The recent trends in sentiment analysis consisting of lexical, machine learning, deep learning, and transfer learning approaches are explained.
- A summary of the applications and challenges of sentiment analysis are discussed to keep up with current trending research.
- An overview of sentiment analysis resources and evaluation parameters are also reported.

In the recent past, there has been an emerging perception that the problem of sentiment analysis is merely a text categorization task that classifies the sentiment into positive, negative, and neutral categories. Recent

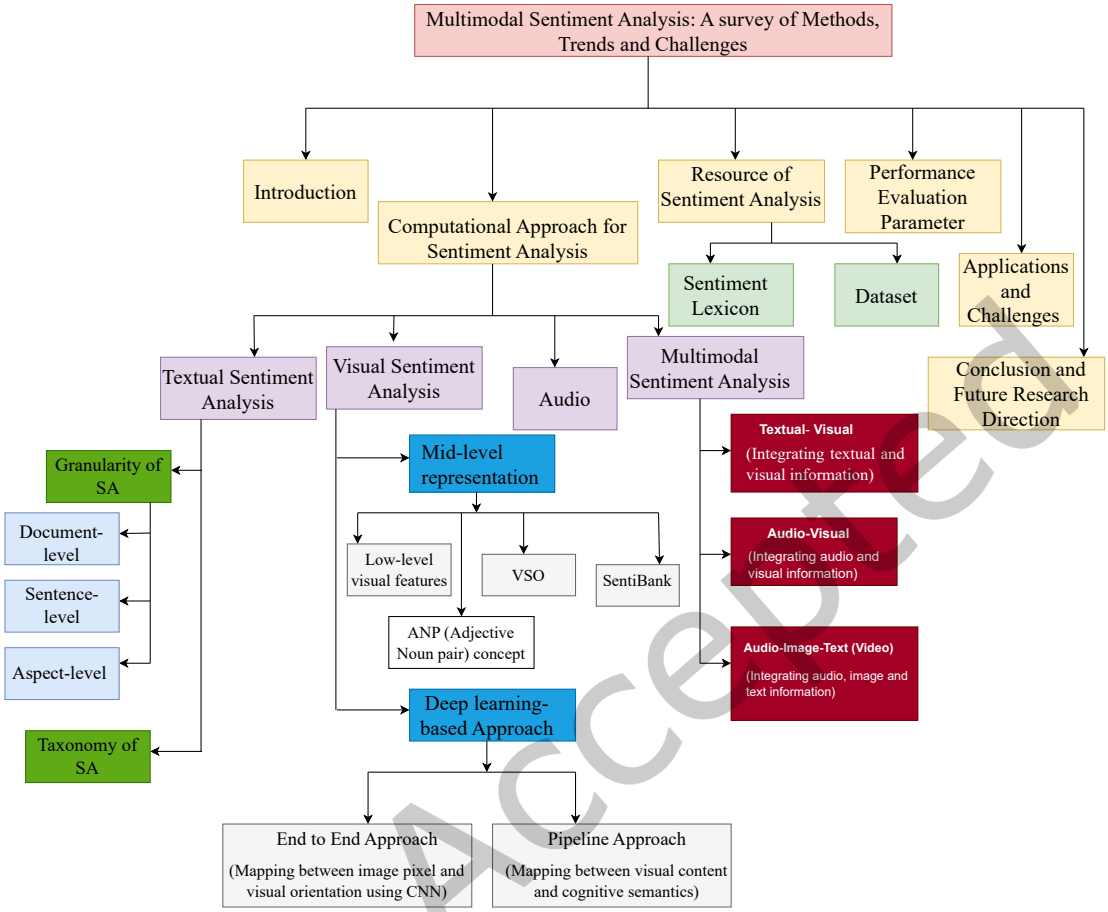

Fig. 2. Organization of the review article

studies have attempted to recognize sentiment expressed in multimedia through multimodal signals, such as visual, audio and textual information. The organization of the review study is shown in figure 2.

2 TEXTUAL SENTIMENT ANALYSIS

Text sentiment analysis can be classified into three granularities: document-level, sentence-level and aspect-level sentiment classification.

2.1 Document-Level

Document-level sentiment classification takes the whole document as a primary unit of information focusing on one topic or object. It is further categorized into positive polarity or negative polarity. Pang et al. [5] are the pioneers of the machine learning-based sentiment classification approach using Naïve Bayes, support vector machines and maximum entropy classifier on the IMDB movie reviews dataset. They experimented with various feature engineering, where the SVM classifier with unigram features yielded the highest accuracy of 82.9%. They claimed that focus detection, discourse analysis and coherence resolution could improve accuracy.

Das et al. [22] performed opinion polarity classification on the Bengali news dataset using a support vector machine classifier from the subjective portion of the sentence. Besides machine learning classifiers, linguistic syntactic features and lexicon entities were used to make the model hybrid. Their system identify the semantic orientation of an opinionated phrase as either positive or negative sentiment polarity and achieved a 70.04% precision and 63.02% recall score.

Similarly, a document-level sentiment analysis framework for the Manipuri language was studied by Nongmeikapam et al. [23]. They collected the data from letters to the editor in local newspapers for the experiment. Text pre-processing was performed using part of speech tagger (POS) and then identified the verbs utilizing a modified verb lexicon and the conditional random field (CRF). The modified verb lexicon was used to predict the positive, negative, and neutral sentiment polarity and achieved a precision of 78.1%, a recall of 72.1% and an F-measure of 75%.

Das et al. [24] presented an Assamese news document sentiment analysis framework using machine learning algorithms and lexical features. They gathered the news manually from various local Assamese newspapers. The random forest classifier with the additional combination of an adjective, adverb and verb lexical features achieved the highest accuracy of 67% among other classifiers. The inadequate feature set and ambiguity in functional words are the limitations of their reported model.

2.1.1 Multilingual Sentiment Analysis. Multilingual sentiment analysis techniques have been developed with the aim of analyzing data in different languages. One of the main problems in multilingual sentiment analysis is the significant lack of resources [25]. Due to the lack of linguistic resources, sentiment analysis in multiple languages is addressed by transferring knowledge from resource-rich to resource-poor languages. Alternate approach is to use a machine translation system to translate texts in other languages into English [26]. Translation systems, however, have various problems, such as sparseness and noise in the data. Sometimes the translation system does not translate essential parts of a text which can cause serious problems.

Mihalcea et al. [27] explored the lexicon and corpus-based methodology for multilingual subjectivity detection. They used the OpinionFinder [28] for English terms lemmatization and translated it into Roman terms. A Naïve Bayes classifier was employed on the Romanian training dataset and achieved F-score of 67.85. Denecke et al. [29] also introduced a text polarity prediction framework within a multilingual framework by leveraging SentiWordNet [30] lexical resource. First, they translated the document into a different language using standard translation software. The author compared the statistical polarity classifier to a method based on n-grams on German movie reviews from Amazon for multilingual polarity detection.

Due to the multilingual nature of social media data, analysis based on a single official language may carry the risk of not capturing the overall sentiment of online content [31]. In order to capture the sentiment from the informal languages, the multilingual sentiment analysis concept plays an important role. The common techniques used are the automatic translation of the target language into English and the subsequent use of the methods and resources available to English or the use of parallel corpora. Balahur et al. [32] developed a simple sentiment analysis system for tweets in English. Subsequently, they translated the data from English to four other languages - Italian, Spanish, French and German using the machine translation system. It is found that joint training datasets from languages with similar structures help to achieve improvement over the results obtained on an individual language. Cui et al. [33] also focused on the multilingual sentiment analysis problem without using the machine translation approach. They analyzed the emotion tokens or SentiLexicon including emoticons, combined punctuations and irregular forms of words. A generative cross-lingual mixture model (CLMM) is proposed by Meng et al. [34] to leverage unlabeled bilingual parallel data. From the experiment results, it is observed that multilingual sentiment analysis using a parallel corpus instead of machine translation can improve classification accuracy.

Xui et al. [35] introduced the instance-level transfer learning scheme applied to cross-lingual sentiment analysis. They translated the important markup languages into the target language as additional training data to enhance the sentiment classification in the target language. TrAdaBoost [36] algorithm was employed to reduce the impact of a low-quality translation corpus. The algorithm effectively improved the sentiment analysis of resource-constrained languages by using large cross-lingual and small target language training data.

A comprehensive study on multilingual sentiment analysis has been done by Lo et al. [37]. The author reviewed the different approaches and tools used for multilingual sentiment analysis, identifies challenges and provides several recommendations including a framework that is particularly applicable to dealing with resource-poor languages. Dashtipour et al. [31] discussed existing approaches for multilingual sentiment analysis and compared several of them on the same datasets. Again, Araújo et al. [38] evaluated sentence-level sentiment polarity classification methods proposed for English and specific for other languages. A survey on the approaches to each of the tasks of sentiment analysis including multilingual technique, as well as supporting language resources, tools, lexicons, corpora, ontologies and datasets in the context of Portuguese was reported by Pereira et al. [39].

2.2 Sentence-Level

A sentence-level sentiment classification restricts the analysis to individual sentences. Hu et al. [40] proposed a lexical-based approach for review mining and summarization for product review datasets. Their research focused on generating feature-based summaries of online customer reviews. First, they extracted the product features that the customer had reviewed. The next step is identifying the opinion sentence as positive and negative polarity and summarizing the results. Finally, predict the sentiment from the product review sentences. Though the model showed a good result, it can not handle the pronoun resolution problem.

Hasan et al. [41] adopted the machine learning approach with unigram features for binary sentiment analysis on the Twitter dataset. Polarity and subjectivity were calculated using SentiWordNet, W-WSD and TextBlob libraries. Unigram features with the Naïve Bayes classifier achieved the highest accuracy of 79% for binary classification. They have also translated the Urdu Tweets into English for classification. They found that the N-gram features with an SVM classifier became the most effective model.

A weakly-supervised deep embedding (WDE) sentiment analysis framework was proposed by Zhao et al. [42] on the review rating dataset to train the system. They used a convolutional neural network to construct WDE-CNN and LSTM for making WDE-LSTM which extracts feature vectors from rating review sentences. The proposed system was evaluated on an Amazon dataset from three domains: cell phones, digital cameras and laptops. The accuracy obtained on the WDE-LSTM model was 87.9%, and on the WDE-CNN model was 87.7%. The experiment results depicted that deep learning models give the highest accuracy compared to baseline models especially when the input is very information rich.

2.3 Aspect-Level

Aspect-level sentiment analysis is also known as feature-based or entity-based sentiment analysis. It includes identifying features or aspects in a sentence and categorizing them as positive or negative. An aspect may be explicit or implicit in the text. Explicit aspects are found directly in the sentence. Implicit aspects are not explicit in the sentence, however, can be inferred from the expressions of sentiment. Aspect-based sentiment analysis is comprised of aspect extraction and identification, the grouping of aspects, sentiment classification and aspects summarization [39].

An implicit aspect extraction and polarity detection framework was suggested by Xu et al. [43] on the Chinese reviews dataset using an explicit topic model. They extend a popular topic modeling method, called Latent Dirichlet Allocation (LDA), to construct an explicit topic model. Before applying topic modeling, a few explicit aspects were assigned to different topics using LDA. The explicit features were extracted by word segmentation,

POS tagging and feature clustering. Feature grouping is performed on the words and phrases based on the same domain features. Then the word clusters are chosen as the training attributes for the classifiers using a support vector machine classifier. The topic model algorithm with the SVM classifier yielded an F-measure of 77.78% for implicit and explicit features classification.

An aspect-level sentiment detection framework named Sent_Comp was developed by Che et al. [44] using the sentence compression technique. Syntactic and extractive compression techniques were used for sentence compression which removes unnecessary sentiment information. They employed a discriminative conditional random field model to compress the sentence information automatically. Though Sent_Comp is domain-independent, information loss is the main issue for a compression technique. They found a strong relationship between aspects and polarity words in aspect-based sentiment detection which greatly affects the efficiency.

Akhtar et al. [45] proposed a multitask learning framework for the identification and classification of aspect terms in a unified model. They utilized a Bi-LSTM network followed by a self-attention mechanism for aspect sentiment classification. The proposed approach employed a BiLSTM followed by a self-attention mechanism to identify the aspect terms in a given sentence. After that, the architecture utilizes a CNN framework to predict the sentiments of the identified aspect terms. A common limitation of these studies was that they failed to prune aspects which resulted in many incorrect aspects. Attention with LSTM mechanisms was utilized by Wang et al. [46] for aspect-based sentiment analysis to focus on the different aspects of the sentence. The attention mechanism was used to concentrate on different parts of a sentence when different aspects were taken as input. The model was trained in an end-to-end way by backpropagation, where the loss function was the cross-entropy loss.

3 TEXTUAL SENTIMENT ANALYSIS APPROACHES

The field of text sentiment analysis has started as an alternative way to topic detecting that aims to extract evaluative meaning. Sentiment analysis is classified into document-level, sentence-level and aspect-level granularities (shown in figure 2). To solve different levels of sentiment analysis tasks, lexical, machine learning and deep learning methods are widely used.

3.1 Lexicon-Based Approach

The lexicon-based approach is further classified into the dictionary-based and corpus-based approaches. The dictionary approach aims to develop a word dictionary consisting of synonyms and antonyms for each word. On the hand, the corpus-based is a data-driven approach by accessing the sentiment labels as well as the context of the document. It uses a corpus integrated with linguistic pre-set assumptions and explanations.

3.1.1 Dictionary-Based Approach. Kim et al. [47] presented a dictionary-based approach to determine the sentiment of opinions. They collected a small set of opinion words with their sentiment orientation. After that, include the synonyms and antonyms of the corresponding opinion words with the help of WordNet [48] and thesaurus [49]. After the newly discovered words are added to the seed list, the next iteration begins. When no more words are found, the iterative process stops. Manual inspection can be performed to remove or correct errors after the process is complete.

Taboada et al. [50] extended the semantic orientation calculator (SO-CAL) dictionary to predict the sentiment polarity. SO-CAL is a dictionary of annotated words and their semantic orientation with positive-negative intensity on blog postings and video game reviews. It was used to measure the subjectivity and opinion in a text. The semantic orientation was computed using a simple aggregate-and-average method where the total score of all adjectives was divided by the total number of adjectives in the document.

Apart from English, a Hindi movie review sentiment classification framework was reported by Pandey et al. [51] by using the Synset replacement algorithm and HindiSentiWordNet (HSWN). Words that are not found

in HSWN are replaced with the same meaning word that is present in HSWN. The review polarity words are extracted using the Hindi lexicon which is finally aggregated to calculate the positive, negative and neutral sentiments. They also resolve the negation and discourse relations to improve the system's performance.

Qiu et al. [52] proposed an advertising strategy named Dissatisfaction-Oriented Advertising based on sentiment analysis (DASA), to simultaneously improve ad relevance and user experience. Specifically, by using syntactic parsing and a sentiment dictionary, they proposed a rule-based approach to extract topic advertising keywords of opinion sentences associated with negative sentiment. The experiment results demonstrated the effectiveness of the proposed approach on ad selection and advertising keyword extraction.

3.1.2 Corpus-Based Approach. Nasukawa et al. [3] created a sentiment lexicon with 3513 sentiment terms to determine subject favorability. It examined how the subject term modifiers and phrases interact syntactically. Their approach extracted the sentiment associated with specific subjects from a document rather than classifying the whole document as positive or negative. They focused on identifying semantic relationships between sentiment expressions and subject terms to achieve a high precision rate.

A set of patterns of tags was utilized by Turney et al. [6] for extracting two-word phrases from the reviews dataset. They used a part of speech (POS) tagger to identify the adjective and adverbs from the reviews. To determine the semantic orientation of the review, the PMI-IR algorithm was employed by issuing queries to a search engine, where "poor" and "excellent " are considered the boundary polarity orientation words. Further enhancement can be made by tagging sentences based on the whole or partial element of the discussion.

Lu et al. [53] presented an approach to estimating the sentiment strength of user reviews according to the strength of adverbs and adjectives expressed by users in their opinion phrases by the progressive relation rules. They multiplied the strength of adverbs and adjectives to handle the sentiment intensity problem. Experimental results on a hotel review dataset in Chinese revealed that the proposed approach was effective in sentiment classification and achieved a good performance on a multi-scale evaluation.

3.2 Machine Learning-Based Sentiment Classification

A machine learning-based sentiment classifier aims to construct predictive models on annotated datasets from which it can learn automatically. This approach generates a feature vector of each text document in which certain aspects or word frequencies are analyzed to train the model and then validate them against reference annotated text documents [54]. Figure 3 shows the working principle of the machine learning-based sentiment analysis model.

Pang et al. [55] utilized the cut-based approach for subjectivity classification on the movie review dataset. The cut-based subjectivity detectors determine the subjectivity portion from all document sentences using pair-wise relationship and per-item information. They first extracted the subjective portion from the review document and classified it into subjective and objective categories using a minimum-cut framework. The experiments showed that utilizing a minimum-cut framework improved the impact of sentiment analysis.

A comparative experiment was performed by Cui et al. [56] with different machine learning classifiers on the product reviews dataset. A generative model, winnow [57] and a discriminative algorithm was employed for classification. They showed that NB performed better than SVM using the same unigram features on datasets other than movie reviews. Also, it was revealed from the experiment result that the high-order n-grams helped discriminate the articles' polarity in the mixture context. But a large number of n-grams are likely to be noisy and a better feature selection approach is needed.

Niu et al. [58] addressed a sentiment analysis model in the medical area using the support vector machine classifier and language model. To extract the polarity information author used the medical outcomes of patients' records instead of identifying their personal opinion. They manually gathered a small set of medical domain

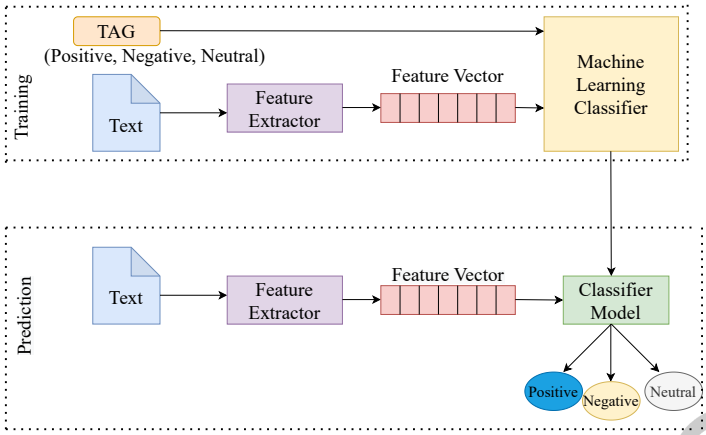

Fig. 3. Machine learning-based sentiment analysis

lexicon words. Apart from the language model, domain knowledge and linguistic features were incorporated to improve the system's performance.

A comparison between support vector machine and Naïve Bayes sentiment classifiers was performed by Zhang et al. [59] for a review dataset written in Cantonese. Unigram, unigram_freq, bigram, bigram_freq, trigram, and trigram_freq were used for representing the text. They studied how feature size and feature representation could affect classification performance. The results revealed that the NB classifier performed better than SVM with 95.67% for 900–1100 features. Singh et al. [60] also used lexicon-based, machine learning and deep learning methodologies to compare sentiment analysis on the Manipuri video review comments. A total of 4000 Manipuri review comment dataset was manually categorized with positive and negative sentiments after collecting from social media platforms. The Naïve Bayes (NB) classifier performed better than other classifiers. Again, a Manipuri news sentiment analysis system using the machine learning approach was introduced by Meetei et al. [61]. They collected the data from the local daily newspaper. The authors carried out some language-specific pre-processing tasks like transliteration, building a negative morpheme-based lexicon and filtering noisy words to make the system more robust. Using additional linguistic features in their models achieved the highest accuracy.

3.3 Hybrid Sentiment Classification

A concept-level sentiment analysis framework was reported by Poria et al. [62] on the movie review and the Blitzer dataset. The authors combined linguistics and common-sense computing with a machine learning classifier to train the model. They also combined the sentic computing framework with dependency-based rules to achieve better polarity detection and understanding of the contextual role of a sentence. They found that the extreme learning machines (ELM) performed better than the support vector machine classifier in training time and accuracy.

Appel et al. [63] investigated a hybrid sentiment analysis approach using SentiWordNet and machine learning classifiers on the movie review and Twitter dataset. They used the sentiment lexicon enhanced with the assistance of SentiWordNet and fuzzy sets to estimate the semantic orientation polarity and its intensity for sentences. The use of argot, jargon, idiom, and lingo is hard to deal with, and sometimes it misguides the system in detecting opinions properly and is a challenge for the system.

Apart from review comments, Ortigosa et al. [64] suggested a sentiment classification and change detection model using a lexicon and machine learning approach on the Facebook comments dataset. The sentiment lexicon

model was developed based on word count (LIWC), Spanish linguistic inquiry and slang comments from Facebook. Naïve Bayes and support vector machine classifiers were employed to evaluate the developed lexicon and yielded 83.27% and 83.27% accuracy.

3.4 Deep Learning-Based Sentiment Classification

The deep learning-based sentiment analysis model has become very popular due to its outstanding performance. Deep learning is an emerging machine learning area that offers supervised and unsupervised feature representation approaches [65]. This approach contains multiple perceptron layers inspired by the human brain [66]. As a part of the machine learning approach, Hinton et al. [67] first proposed a deep brief neural network in 2006. One of the primary deep learning models, Recursive Neural Tensor Network (RNTN), was introduced by Socher et al. [68] for sentence-level sentiment classification by modeling the effect of the composition of the phrases. They also proposed the Stanford sentiment treebank corpus consisting of parse trees with the sentiment. The pre-trained network can perform better after fine-tuning the model rather than simply using the feature vector that fails to capture the discriminative knowledge [69].

Kim et al. [70] adopted a convolutional neural network architecture to deal with the textual sentiment classification. The proposed system was a combination of an embedding layer, two convolutional layers, one pooling layer and one fully connected layer. They used customer reviews, movie reviews and Stanford sentiment treebank datasets to compare conventional machine learning classifiers and deep neural architecture. Their proposed model showed that parallel convolutional neural network architecture reduced computational overhead and improved classification accuracy. They concluded that using consecutive convolutional layers is effective for relatively long texts.

Researchers are integrating different deep-learning techniques for sentiment classification. A deep sentiment classification framework was explored by Huang et al. [71] by combining different deep learning architectures. The proposed system consists of pre-trained word vectors as input and employs one layer of convolutional neural network and the layers of long short-term memory stacked on CNN. CNN is used to gain significant local text features and LSTM can extract context-dependent features and generate sentence representations for sentiment prediction. The experimental results demonstrated that the model outperformed existing LSTM, CNN, CNN-LSTM and SVM classifiers.

Rhanoui et al. [72] reported a document-level sentiment analysis framework using a deep convolutional neural network and bi-directional long short-term memory on French news articles. The CNN extracts the local features, and the extracted features are fed as the input to long short-term memory. Along with four filters, one max pooling layer was applied after each filter. The results of max pooling were concatenated and passed as the input to b-LSTM. Again, the output of b-LSTM was passed as the input to a fully connected layer. Finally, the softmax function was applied as an activation function to produce the desired output by assigning classes to articles. Experimental results revealed that the CNN-BiLSTM and Doc2Vec embedding achieved high results for sequence classification tasks on long texts.

Shalini et al. [73] worked on a Bengali-English code-mixed language for sentiment classification. They applied a convolutional neural network architecture to the code-mixed data, i.e., the SAIL_ICON_2017 dataset. The authors also performed the same experiment on the monolingual language, i.e., Telugu online movie reviews dataset using the same deep neural architecture. The code-mixed Bengali data yielded an accuracy of 0.732 and 0.513 accuracy achieved on the Telegu dataset. Telugu is morphologically rich and has a higher chance of words with the same meaning. Therefore, the cross-validation accuracy for the Telugu dataset was lower. Using the deep ensemble model, a code-mixed sentiment analysis framework was also reported by Baroi et al. [74] on Hindi and English languages on the social media platform. They used CNN and LSTM-based ensemble models to detect

Hinglish hate speech on the Sem-Eval 2020 dataset. The proposed framework reported an F-Score of 0.617 on the test data.

Researchers begin to experiment with an attention mechanism to improve the system's performance. Yang et al. [75] reported a document-level sentiment classification model using a hierarchical attention mechanism. The vital content of the document was taken into consideration for document representation. The experimental results on the text-based reviews dataset showed that the proposed model outperformed the existing model by capturing insights about the structure of the document. Generating the semantic relations between sentences in a document is a significant challenge in document-level sentiment classification. This problem was addressed by Rao et al. [76] by developing the SR-LSTM model. The first layer of the model was used to extract the sentence vectors using LSTM. The next layer was used to recognize the relationship between sentences.

3.4.1 Transfer learning. Transfer learning is an advanced form of artificial intelligence in which a trained model uses its knowledge to transfer it to a new one. It is an emerging machine learning technique that uses existing knowledge to solve different domain problems and produces prediction results. Nowadays, the transfer learning technique is applied by many scholars to the field of sentiment analysis [77]. Rather than requiring detailed training data, the new model uses the previously learned features directly. Knowledge of one domain can be transferred to another using this technique. Due to its great result and accuracy and requiring less training time, the transfer learning technique has grown exponentially [78].

Yang et al. [79] reported a multi-task learning-based financial aspect-based sentiment analysis framework using ULMFiT [80] which is an inductive transfer-learning method. The proposed system outperformed the traditional transfer learning approach. Even though a modified ULMFiT technique for aspect-based sentiment analysis is trained on a relatively small sample but it showed better system performance. Additional study is required to apply this approach to multi-task and multilabel learning.

Hoang et al. [81] utilized the contextual pre-trained word representations model like BERT combined with a fine-tuning approach with additional generated text for solving out-of-domain aspect-based sentiment analysis problems. They employed the BERT model for pre-training and fine-tuned it to find semantic similarities between a text and an aspect. Though the proposed aspect classifier model could deal with related and unrelated label data, analyzing the review language is a difficult task that requires a deep understanding of that language. This problem was addressed by Karimi et al. [82] by using the hierarchical BERT approach. They proposed hierarchical and parallel aggregation on top of BERT for aspect extraction and aspect sentiment classification.

Peters et al. [83] explored the Embeddings from Language Models (EMLo) method for contextualized word representations. They prepared a deep contextualized word representation approach to deal with the complex word characters, e.g., syntax and semantics and different cross-linguistic contexts like polysemy. The rich word representation can be used for different natural language processing tasks. They first trained a bidirectional language model (BiLM) on the corpus and the output of the BiLM is fed into the long short-term memory network to generate the characterization of words. They reveal that the BiLM layers effectively encode various types of semantic and syntactic information of words in different contexts which can improve the overall task performance. A summary of some of the selective approaches of textual sentiment analysis is shown in Table 1.

3.5 Discussion

Sentiment analysis can be classified into three conceptually distinct methods such as lexicon-based, machine-learning and deep-learning approaches. These groups reflect two major methodological transitions: the first one is from manual labor-intensive and hand-crafted lexicons to an automatic machine learning approach trained on high-dimensional, sparse bag-of-words features and the second one is learning the low-dimensional and dense embeddings vector from texts through artificial neural networks using pre-training based on large-scale open-domain text [84]. From the existing research, it can be observed that machine-learning approaches are

most commonly used in traditional sentiment analysis. Besides traditional machine learning methods, ensemble classifiers are adopted to obtain more precise results. However, the limitations of machine learning methods are mainly reflected in two aspects. One is that the performance of machine learning methods depends on the number of annotated samples. The other is that they are domain dependent. Lexicon-based methods are therefore utilized because they do not require annotated data. However, lexicons are context-independent and these methods depend on static lists of words. Therefore, hybrid approaches are proposed to compensate for the shortcomings of machine-learning and lexicon-based approaches. In addition, deep learning methods have received more attention. Most of the existing sentiment analysis research works rely on deep neural architecture with context-independent word embedding including Word2Vec, GloVe and FastText where there is fixed text representation irrespective of their context. In contrast to the traditional machine learning approach, the more recent transfer learning models can understand the generic expressions and word relationships from related domains and tasks of a given sentiment dataset. Due to their architecture, these methods can learn fundamental text representations that are useful across domains. These transfer learning models are predominantly based on the language model architecture of BERT, a pre-trained context-independent language model revolutionized for natural language processing tasks.

Table 1. Summary of selective approaches of textual sentiment analysis

References	Dataset/ Domain	Method	Performance	Language
Turney et al. [6]	Multi-Domain Review Dataset	Developed a system for rating the thumbs up and thumbs down using PMI-IR and POS Tagger	74% Accuracy	English
Lu et al. [53]	Chinese Hotel Reviews Dataset	Developed a system for determining the sentiment polarity intensity using lexical features	71.6% Precision	English
Pang et al. [5]	IMDB Movie Reviews Dataset	Developed a sentiment classification system using Naïve Bayes, Support Vector Machines, Maximum Entropy	82.9% Accuracy	English
Das et al. [24]	News Domain Dataset	Developed a system for news sentiment classification using machine learning classifiers with lexical features	67.7% F1-score	Assamese
Poria et al. [62]	Movie Review and Blitzer Dataset	A concept-level sentiment analysis framework using linguistics and common-sense computing	86.2% Accuracy	English
Kim et al. [70]	Reviews dataset and Stanford Sentiment Treebank Dataset	Developed a sentiment classification framework using convolutional neural networks	81% F1-score	English
Baroi et al. [74]	SemEval 2020 Dataset	A sentiment classification framework on code mixed language using CNN and LSTM ensemble architecture	61.7% F1-score	Hinglish
Huang et al. [71]	Sina Micro-Blog Dataset	Developed a deep sentiment classification framework by combining one layer of CNN and two-layer of LSTM	87.2% Accuracy	Chinese
Akhtar et al. [45]	SemEval 2014 dataset	A multi-task sentiment classification framework utilized a Bi-LSTM network followed by a self-attention mechanism	83.3% F1-score	English, Hindi
Meetei et al. [61]	News articles Dataset	Developed a sentiment classifier using machine learning and deep learning algorithms	75.5% Accuracy	Manipuri

4 VISUAL SENTIMENT ANALYSIS

Visual sentiment analysis intends to extract sentiment-related information from the visual content. It is becoming a challenging task because of the abstract representation of an image.

4.1 Mid-Level Representation Approach

Attribute learning and mid-level feature representations like ObjectBank [85] began to rise in popularity in the computer vision field. Consequently, researchers strive to bridge the semantic gap between low-level visual features and sentiment orientation using mid-level representations. Thus, various mid-level sentiment representations are reported for visual sentiment classification. A proxy representation in vision systems with higher fidelity could also aid visual sentiment analysis with a mid-level representation approach. Chen et al. [86] reported an image sentiment classifier using the object-based concept and semantic information. Due to the powerful correlation between adjectives and nouns, the authors considered the adjective-noun pairs (ANP) from the image tag. They detected the objects and their corresponding attributes. Experimental results show that the mid-level perspective approach improves sentiment classification performance but increases computational complexity.

Yuan et al. [87] proposed an image sentiment prediction framework named "Sentribute" which consists of 102 mid-level features that leverage the mid-level attributes of an image to determine its sentiment. The mid-level features made the sentiment prediction outcome more meaningful than directly using the low-level image features. They employed the scene-based attribute to represent the mid-level feature. The mid-level features such as the facial expression recognition helps the result of the classification model to be more interpretable than using only the low-level attributes.

Borth et al. [88] developed a large-scale semantic concept visual sentiment ontology (VSO) named SentiBank [89] based on web mining and psychological theories. The authors employed the adjective-noun pair (ANPs) concept to build a large-scale VSO as mid-level descriptors. The adjective-noun tags like 'beautiful flowers' or 'sad eyes' are extracted from Youtube videos and Flickr images. A well-known psychological model of human emotions named as the Plutchik Wheel of Emotion [90] has been used to search for these images and videos. The experiments revealed that the machine learning classifiers such as linear SVM and logistic regression with the SentiBank concept performed better than the text-based approach in tweet sentiment prediction.

A visual sentiment topic model (VSTM) was reported by Cao et al. [91] for image sentiment analysis on microblog images. The usefulness of the VSTM model is that it includes a macro explanation of the visual content of one topic. They gathered images from the same microblog topic to enhance the sentiment results. The authors used visual sentiment ontology (VSO) to extract the visual features. According to the topic model, the visual ontology features were extracted for better system accuracy.

It is hard to identify whether the ANP is highly correlated with the sentiment orientation of visual content using VSO-based models. Therefore, Li et al. [92] presented a visual sentiment analysis approach using the text sentiment information of the ANPs. The author calculated the overall sentiment value of an image according to the text sentiment values of ANPs and the corresponding responses of the particular image. The visual sentiment value is then used as a one-dimensional feature for image sentiment prediction. Adding textual sentiment analysis to image sentiment analysis improved the performance of image sentiment analysis. As a result of their excellent openness and versatility, the VSO and SentiBank are used in a wide range of applications including the viewer emotion prediction and Animated GIFs [93].

Most of the existing mid-level-based approaches list the concepts together to form a mid-level sentiment though it ignores the distinction and connection between the ontology concepts. Because of the remarkable successes achieved by deep learning in computer vision, it has prompted to apply deep learning technologies to visual sentiment analysis. Ortis et al. [94] jointly utilized the objective text description of images extracted from the visual content of images for sentiment prediction. They proposed to extract and employ an objective text description of images rather than the classic subjective text provided by the users. The objective text was obtained through deep learning architectures which were used to classify objects and the scenes to perform image captioning.

4.2 Deep Learning-Based Approach

Recently, deep learning has gained huge attention due to its extraordinary success in artificial intelligence. It uses multi-layer models to transform low-level features into an abstract feature space which can better describe the essential information of input image [95]. This brings new opportunities for computer vision, artificial intelligence and emotional semantic analysis for visual content. The deep learning-based image sentiment analysis approach is classified into end-to-end and pipeline-based approaches.

4.2.1 End to End Approach. You et al. [96] reported an end-to-end technique using a fine-tuned convolutional neural network on Tumblr and Twitter image datasets for performing visual sentiment analysis. From the initial training data, they selected the potentially cleaner training samples. Furthermore, they improved the performance of Twitter images by inducing domain transfer with a small number of manually labeled Twitter images. They used a CNN and a progressive CNN (PCNN) to fine-tune the AlexNet model. The PCNN based model outperformed the CNN based model. Thus, the experimental results revealed that the use of convolutional neural networks improved the accuracy of visual sentiment analysis. Jindal et al. [97] also carried out a visual sentiment detection framework after fine-tuning the CNN deep neural network on the Flickr dataset. The pre-trained CNN used the deep architecture of Krizhevsky et al. [98]. The pre-trained VGGNet and ResNet models were used to extract the image feature to detect the image sentiment using CNN. Domain-specific fine-tuning improved the system performance of the CNN model.

Campos et al. [99] also reported a visual sentiment prediction system using fine-tuned convolutional neural networks. They used the CaffeNet CNN architecture [100] which is an AlexNet-styled network that differs from the ILSVRC2012 winning architecture [98] in the sequence of the normalization and pooling layers. They demonstrated that deep architectures could learn useful features for recognizing visual sentiment in social Twitter image datasets. When the target dataset is small, the choice of pre-training can make a difference. Further, they presented a sentiment prediction visualization with spatial localization to understand erroneous classifications and learned network representations.

Siersdorfer et al. [101] proposed an image sentiment classification to analyze visual content as positive and negative sentiment. The author considered the bag-of-visual word representation and the color distribution of images and used the SentiWordNet lexicon to extract numerical values for the sentiment from accompanying textual metadata. They studied the relationship between the visual content and the sentiment of images. The SentiWordNet [102] lexicon was used to extract a numeric sentiment score assigned to each image and its text content. The experiment results revealed that there is a strong correlation between sentiment scores extracted from Flickr meta-data (e.g., image title, description and tags provided by the user) and visual features (i.e., SIFT-based bag-of-visual words and local/global RGB histograms).

A deep couple adjective and noun (DCAN) pair convolutional neural network was developed by Wang et al. [103] for image sentiment analysis. To reduce the large intra-class variance, a middle-level sentiment representation is shared by jointly learning an adjective and a noun pair. Secondly, based on the learned sentiment representation, a prediction network was further optimized to deal with the subtle differences which often exists in the fine-grained image categories. The three networks were trained in an end-to-end manner for image sentiment classification.

Li et al. [104] proposed a visual sentiment prediction framework by translating images into textual descriptions and analyzing visual sentiment using textual sentiment analysis indirectly. A deep learning-based image caption framework consisting of a deep residual network and a long and short-term memory network was utilized to generate the initial textual description of images. The adjective-noun pairs were added to the textual description obtained by the images caption model to generate the final textual description of images. Then, the text used to describe visual content was processed by means of deleting redundant symbols, retaining part of the vocabulary and embedding word vectors. The generated word vectors were fed into a time series network to train a sentiment prediction model. The network model of image sentiment analysis is shown in Figure 4.

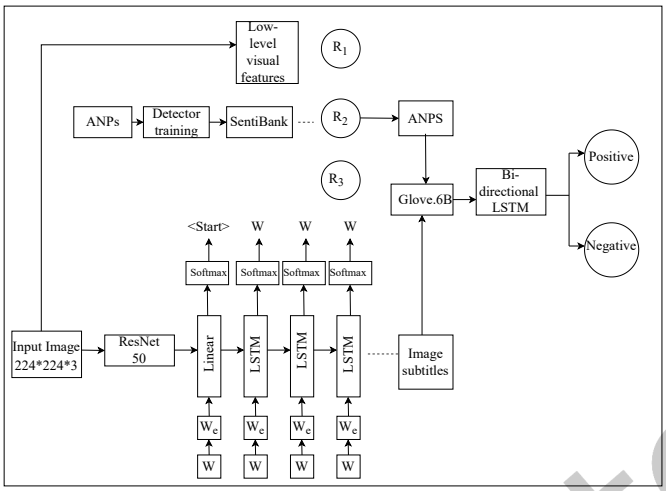

Fig. 4. The network model of image sentiment analysis

4.2.2 Pipeline-Based Approach. A pipeline mode-based visual sentiment analysis uses deep learning models to establish a mapping between visual content and cognitive semantics to predict the sentiment orientation of visual content. Jou et al. [105] presented an ontology with extended breadth and volume in a multilingual visual sentiment ontology (MVSO) consisting of 15,630 ANPs from 12 major languages and 7.37 M images from over 235 countries that considers the cultural differences. MVSO used a similar ANP mining process to VSO focusing on Flickr with more stages of candidate filtering including diversity of users for a given ANP, semantic correctness and language-specific syntax. In addition, the author also explored a method for generating ontology structures automatically by either grouping ANP or by exact translations with a pivot language using word embeddings. These multilingual ANPs and their associated detector banks have applications in portrait analysis, cross-lingual visual sentiment analysis and image query expansion diversification.

Liu et al. [106] explored a cross-lingual image sentiment analyzer using a culturally-coherent image query expansion engine based on MVSO ontology. Cross-lingual image-based sentiment analyzer, an interactive multilingual ontology browser and a culturally coherent sentiment-aware image query expansion engine were the three major functions implemented by their system.

Recently, Ahsan et al. [107] investigated a pipeline-based approach to predict the visual sentiment on social media images. They first generated a series of social event concepts and computed corresponding concept scores using the CNN architecture. This study inspired image sentiment analysis by focusing primarily on social media events. The model could recognize positive images with better accuracy where strong visual cues were presented in the image but made errors when differentiating among positive, negative and neutral sentiments.

Mathews et al. [108] explored a method known as SentiCap that generates image descriptions with positive and negative sentiment polarity. Instead of using only the RNN model, they introduced a novel switching RNN model that combined CNN and RNN models running in parallel. The model consists of two parallel RNNs such that one represents a general background language model; another specialises in descriptions with sentiments. They designed a word-level regularizer so as to emphasize the sentiment words during training and to optimally combine the two RNN streams. In each time step, this switching model generated the probability of switching between two RNNs. The proposed SentiCap model generates the image captions and gives the sentiments with

higher accuracy. The model was a specialized word-level supervision scheme to utilize a small amount of training data with sentiments effectively.

An attention-based visual sentiment analysis framework was adopted by You et al. [109] from the local image regions which can jointly discover the relevant local image regions and build a sentiment classifier on top of these local regions. They studied how the local region of one image is relevant to capturing sentiment. A visual attribute detector model extracted the visual attributes which were fed into one attention module for classification. The proposed model performed better than the state-of-the-art system. Song et al. [110] presented visual sentiment analysis framework named Sentiment Networks with visual Attention (SentiNet-A) by integrating visual attention with a convolutional neural network. To model visual attention, they developed multiple layers to generate the attention distribution over the regions of the image. A summary of the selective sentiment analysis on visual sentiment analysis is shown in Table 2.

Table 2. Summary of selective approaches of visual sentiment analysis

Author	**Domain/ Dataset**	**Approach/Finding**	**Performance**
Yuan et al. [87]	SUN Database	Developed a system named "Sentribute" which consists of 102 mid-level features that leverage the mid-level attributes of an image sentiment	82.35 %
Cao et al. [91]	Microblog Dataset	A Visual Sentiment Topic Model from the icroblog topic images by applying visual sentiment ontology (VSO).	58.9 %
Li et al. [92]	VSO, SentiBank1.1, Flickr Dataset	Presented a visual sentiment analysis approach by using the text sentiment information of the ANPs.	87.6 %
You et al. [96]	Twitter and Flickr Dataset	Design framework using a CNN architecture and progressive strategy to fine-tune the deep network for image sentiment analysis.	87.6 %
Mathews et al. [108]	MS-COCO Dataset	Explored a method known as SentiCap that generates image descriptions with positive and negative sentiment polarity	88.4 %
Campos et al. [99]	Twitter Dataset	A visual sentiment prediction system using fine tunning the CNN for local patterns that the network learned to associate with image sentiment	80.3 %
Ortis et al. [94]	Instagram and Flickr Dataset	Utilized the objective text description of images with the CNNs model to perform image captioning and inferred the sentiment polarity	75.9 %

4.3 Discussion

There is progress in sentiment analysis for visual content in social media and other domains. Mid-level representation-based and deep learning-based approaches are the main solutions to bridge the semantic gap between visual content and sentiment orientation. Mid-level representation methods detect the presence of cognitive concepts in visual content and use the responses of these concepts as middle-level features to predict sentiment orientation through machine learning. Deep learning-based methods create a direct mapping between image pixels and sentiment orientation. These deep learning-based methods predict the sentiment of visual content after detecting the presence of ontology concepts using deep learning models. Although these methods have shown to be successful, the complexity of visual data and the unreliability of sentiment labels still affect the performance of these models for visual sentiment prediction.

5 SENTIMENT ANALYSIS USING AUDIO

Audio sentiment analysis is increasingly relevant in sentiment analysis tasks as the number of online videos published on social media platforms has increased over time. However, there is lesser research work on audio sentiment analysis. Automatic speech recognition (ASR) technologies are commonly employed for speech-based sentiment analysis to convert speech into text. Therefore, sentiment prediction depends mainly on textual sentiment analysis and automatic speech recognition performance. Detecting sentiment solely from spoken words is a comparably new field.

Automatic speech recognition (ASR) technologies convert speech into text to predict sentiment from audio. Ezzat et al. [111] explored conventional text mining techniques over transcribed audio recordings to detect the speakers' emotions. They converted speech into text by using automatic speech recognition technology. The authors also study the effects of various methods for selecting features. To detect the speaker's emotions and predict whether the customer was satisfied or not satisfied with the service provided, they developed several text mining techniques based on recorded telephone calls mimicking real agent or customer conversations after translating them into text.

Rather than simply converting speech to text, a single keyword spotting (KWS) system was developed by Kaushik et al. [112], [113] for audio sentiment prediction. To identify the sentiment-bearing terms in KWS, they applied the text sentiment classifier which reduces the complexity of the textual sentiment classification model. Finally, they utilized the sentiment-bearing information to develop a language model which was more specific. The proposed model was evaluated on UT-Opinion and YouTube video corpus. They observed that sentiment analysis on natural speech could be understood clearly even with low word recognition rates.

Recently, Amiriparian et al. [114] compared the conventional frequency cepstral coefficients (MFCC) approach and deep spectrum features for speech sentiment analysis on the YouTube movie reviews dataset. They combined the spectrum features with bag-of-audio-words (BoAW) for feature representation and applied a convolutional neural network for sentiment prediction. The author tested whether it would be suitable for speech sentiment analysis using deep spectrum feature representations.

Abburi et al. [115] also presented a sentiment analysis system from online spoken reviews based on its multimodality nature. They extracted MFCC features only from stressed regions of audio data rather than using them from the entire input. They determined the sentiment of spoken reviews using both textual and audio features. The audio input was transcripted into text and Doc2vec method was used to represent the text feature. A support vector machine was employed to develop a sentiment model on the textual features.

Based on VSO and SentiBank concept, Sagar et al. [116] explored audio content using concept pairs that uniquely combine audio and multimedia information to complement the audio and multimedia applications. They proposed a large-scale folksology AudioSentiBank corpus containing over 1,123 adjective and verb-noun pairs. It is the first time that acoustic concept pairs have been classified which is highly significant in audio sentiment analysis. They further presented a benchmark corpus for acoustic concept pair classification. This is also the first attempt to explore the classification of acoustic concept pairs such as happy crowd and angry crowd.

Many existing works focus on sentiment analysis exclusively from the textual content as present in the speech. Pereira et al. [117] adopted an information retrieval approach for speech sentiment analysis. Specifically, they were looking for documents with opinions similar to the one in the query. Three important contributions were made in this paper. First, they introduced a framework for analyzing sentiments based on polarity which included the ability to accommodate different modalities when none were available. The second finding was that regularization could improve speech transcriptions' sentiment retrieval accuracy. In addition, they demonstrated that their approach was robust by training regularizers on one dataset and then performing sentiment retrieval experiments on another dataset which yielded substantial results. A summary of selective sentiment analysis works based on the audio is shown in Table 3.

Table 3. Summary of selective approaches of audio sentiment analysis

Author/Reference	Dataset	Approach/Finding	Performance
Ezzat et al. [111]	Call center audio Dataset	Designed a Speech to text (traditional) approach audio sentiment detection	92.7 %
Pereira et al. [117]	CNET and YouTube Dataset	Utilized BoW with text classifier for audio sentiment analysis	72.6 %
Kaushik et al. [112]	YouTube and UT-Opinion Dataset	A single keyword spotting system (KWS) is developed for sentiment detection	91.1 %
Abburi et al. [115]	YouTube and MOUD Dataset	Improved multimodal sentiment detection approach by utilizing MFCC features with a GMM classifier	74.7 %
Kaushik et al. [113]	YouTube and UT-Opinion Dataset	A keyword spotting is proposed to determine the sentiment-bearing keyword for sentiment detection	92 %
Amiriparian et al. [114]	YouTube Dataset	A deep spectrum feature representation with CNN was used for performing speech-based sentiment analysis	74.5 %

6 MULTIMODAL SENTIMENT ANALYSIS

With the advances in technology, people increasingly use audio (speech) and visual (videos, images or clips) modalities to express their thoughts. Multimodal sentiment analysis adds an extra flavor to the conventional unimodal approach by including additional visual and audio data information. Most of the multimodal sentiment analysis approaches focus on modality fusion schemes.

6.1 Image-Text approach

A joint image-text sentiment classifier uses a new feature vector obtained by employing a fusion technique from both image and text data to predict sentiment. Wang et al. [118] proposed a joint image-text framework for the microblog images using a cross-media bag of words approach. They collected a total of 5000 microblog images from the Sina Weibo website. After employing a feature-level fusion technique, logistic regression, SVM and naïve Bayes machine learning classifiers were used to train the proposed system. The experimental results reveal that the joint image-text model outperformed the text-based model. Similarly, a fusion-based joint image-text multimodal sentiment analysis model was reported by Chen et al. [119] on the microblog dataset. They presented and visualized the sentiments of microblog data by organizing the results on topic, region and content respectively. The sentiment score was obtained after fusing the sentiment classification scores from both visual and textual channels. The cross-model sentiment analysis was classified into positive and negative polarity corresponding to the microblog, region and topic data. By doing so, social multimedia sentiment can be displayed in a user-friendly and multi-level format.

You et al. [120] carried out a joint textual and visual sentiment analysis framework using deep neural networks. They used a convolutional neural network and a distributed vector model to extract the visual and textual features. Then feature-level, decision-level and cross-modality consistent regression (CCR) fusion methods were employed to execute the joint textual-visual sentiment analysis model. The proposed CCR framework was trained using

deep neural architecture to develop the final sentiment classifier. The experimental results showed that the joint textual-visual model performed better than the unimodal.

An image-text consistency-driven sentiment analysis model was reported by Zhao et al. [121] by exploring the correlation between image and text data on the movie review dataset. A convolutional neural network and long-short term memory are used to decide whether the image and text content are consistent for predicting the sentiment from the joint image-text features. More specifically, the mid-level visual features are extracted by the SentiBank lexicon which is used to represent the visual concepts. After that, they integrated all the feature vector to develop the multimodal sentiment analysis framework. The hybrid multimodal adaptive sentiment analysis model performed better than the deep learning approach.

Li et al. [122] proposed a multi-window convolutional transformer model also known as ConvTransformer which takes advantage of both the Transformer and CNN to detect the joint sentiment polarity. By capturing local n-gram features, the proposed ConvTransformer preserves sequential information of texts and important local n-gram features. Further, they developed a sentiment-aware attention mechanism to be aware of the sentiment intensity information which considered sentiment and positional information to incorporate the intensity of each word by utilizing an external knowledge base (i.e., SentiWordNet).

A multi-stage multimodal sentiment analysis model on the Assamese language was proposed by Das et al. [123] using deep learning architecture that concurrently exploits textual and visual cues to determine the sentiment. The author presented an Assamese multimodal dataset comprising 16,000 articles from the news domain as a benchmark resource. The text model encoded the semantic content by considering the semantic information of the news. To predict the text sentiment polarity, the CNN and LSTM deep learning techniques are employed. After converting the text strings into a vector of numeric values, they are fed into the convolution and global max-pooling layers with long-term dependencies. At the same time, the visual model encoded the visual appearance information from the news image. A news caption generation method for image sentiment prediction was proposed to translate images into a textual description (i.e., image caption or visual description) that is related to the news event. Then, an intermediate fusion-based multimodal framework was employed to exploit the internal correlation between textual and visual features for joint sentiment classification. The intermediate layer of image and text modality was integrated to extract the multimodal cues. The combined features are fed into the fully connected (FC) layers to encode the internal correlation between text and image features. Finally, a decision-level fusion mechanism was used on the three models to integrate cross-modal information effectively for final sentiment prediction. Figure 5 shows the graphical representation of the multi-stage multimodal framework. Das et al. [124] also reported an Assamese image-text sentiment analysis framework using late fusion approach.

Multidomain sentiment classification focuses on transferring information from one domain to the next domain where the first models are trained in the source domain; the knowledge is then transferred and explored in another domain. Yuan et al. [125] proposed a Domain Attention Model (DAM) for modeling the feature-level tasks using attention mechanisms for multidomain sentiment classification. The DAM consists of a domain module and a sentiment module. The domain module was used to predict the domain in which the text belongs using bi-LSTM and the sentiment module was used to select the important features related to the domain using another bi-LSTM with an attention mechanism. To predict the polarity of the sentences, the vector acquired from the sentiment module is fed into a softmax classifier.

You et al. [126] reported a new sentiment analysis framework that combined visual and textual information. They believed that visual and textual information should be treated in a separate structure and joined. A tree-based LSTM model was introduced for aligning visual regions with textual words. The attention mechanism with the LSTM model learned a deeper inside of joint representation for both visual and textual models. They built a semantic tree structure based on sentence parsing to align textual words and image regions for accurate analysis. Next, the system learned a robust joint visual-textual semantic representation by incorporating an attention mechanism with LSTM and an auxiliary semantic learning task.

A combination of visual-textual sentiment framework named multimodal attentive fusion was suggested by Huang et al. [127]. First, two separate attention models were introduced to individually learn the text and image models named the semantic and visual attention frameworks. After that, a multimodal attention model was explored to extract the correlation between textual and visual features. Finally, a decision level was employed to integrate all three attention models to predict the sentiment polarity.

A target-oriented multimodal sentiment classification (TMSC) framework was proposed by Yu et al. [128] by adapting BERT architecture with an attention mechanism to get the target-sensitive text representations. They also presented two extended versions of standard BERT models named as mBERT and TomBERT. The target-oriented multimodal BERT (TomBERT) can effectively capture the inter and intra-modality dynamics. To develop intra-modality dynamics, they first applied BERT to obtain target-sensitive text representations. After that, they borrowed the idea of self-attention and designed a target-attention mechanism to perform target-image matching to derive target-sensitive visual representations. To develop intermodality dynamics, the author further proposed to stack a set of self-attention layers to capture multimodal interactions.

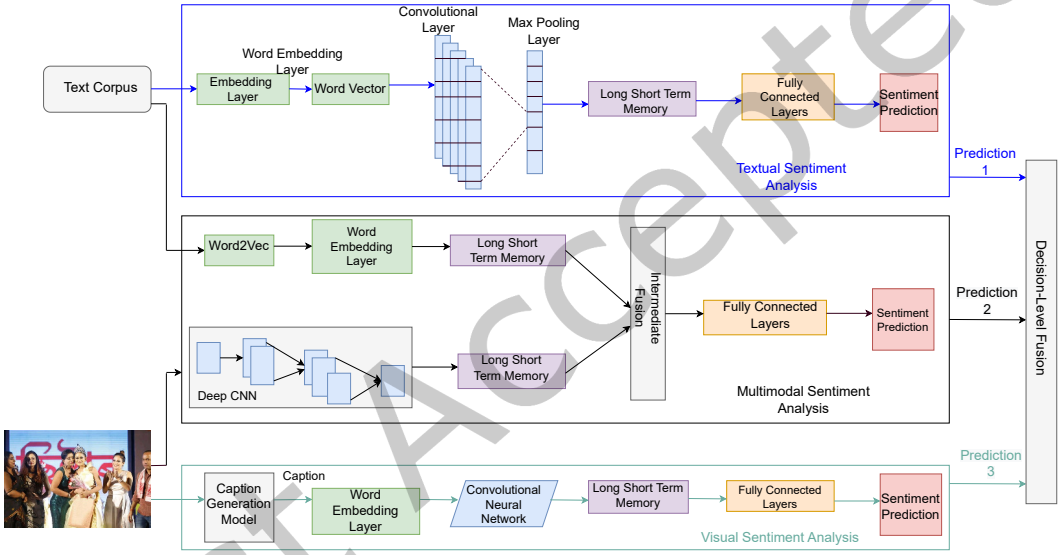

Fig. 5. Multi-stage multimodal framework for image-text sentiment classification

6.2 Audio-Visual approach

The existing work on audio-video multimodal sentiment analysis shows that audiovisual fusion mainly focuses on human faces. An audiovisual sentiment analysis framework was proposed by Yadav et al. [129] by extracting emotion-related information from both the video and the audio channel in audiovisual content. For visual analysis, they extracted the facial expression features. The audio features such as pitch, pause, loudness and voice intensity were extracted from audio data. After integrating the feature vector, they predicted the sentiment polarity of the overall review according to the detected emotions. Chu et al. [130] proposed an audiovisual analysis method to generate emotional arcs for movies, including short videos on the web. They trained audio and video sentiment analysis models and then used them to construct separate emotional arcs for audio and visual content. They also conducted experiments to evaluate the micro-level performance and synthesize the prediction results from audio and visual content.

Jiang et al. [131] proposed a comprehensive sentiment prediction framework from user-generated videos from social media. A kernel-level multimodal fusion was employed on the extracted semantic, audio and low-level visual descriptors features. They presented a dataset collected from user-generated videos with predicted emotions. The dataset consists of 4486 videos from YouTube and 3215 videos from Flickr which were annotated based on the eight emotion categories in Plutchik's wheel. Since more emotion categories were considered, this is somewhat different from sentiment analysis. However, it can be utilized in sentiment analysis because the eight emotion categories can be easily classified into positive and negative sentiment orientations. Ten annotators were asked to filter the videos and 1101 videos were finally kept to build the dataset which can be further used for multimodal research.

A context-aware multimodal sentiment analysis framework for Persian was presented by Dashtipour et al. [132] on the Youtube videos which simultaneously exploits visual, acoustic and textual cues to determine the sentiment with high accuracy. They employed both decision-level and feature-level fusion methods to integrate affective cross-modal information.

6.3 Audio-Image-Text approach

Adding multiple modalities such as audio, image and text makes the sentiment analysis model robust. Morency et al. [133] pioneered multimodal sentiment analysis by proposing a tri-modal sentiment analysis architecture for Spanish. The tri-modal sentiment analysis framework exploited the audio, visual and textual modalities. Further, they identified the subset of audio-visual features relevant to sentiment analysis and presented guidelines for integrating these features. They showed the benefit of visual and audio features and presented the integration guideline for sentiment analysis. The authors also presented a dataset from YouTube.

Cambria et al. [134] proposed a deep neural framework for multimodal sentiment analysis. They extracted the textual features using a convolutional neural network and the OpenSMILE software for audio features. The visual features are extracted using a deep convolutional neural network. All the feature vectors were combined for sentiment classification using support vector machine classifiers.

Poria et al. [135] proposed a multimodal sentiment analysis framework on Youtube videos using multiple kernel learning (MKL) approaches. The short multimodal features were extracted to represent one sentence each. The textual features from uttered sentences and audio data such as vocal pitch and facial expression were extracted using a deep convolutional neural network model. Finally, MKL was used to classify the multimodal heterogeneous fused feature vectors. They presented a parallelizable decision-level data fusion method which is much faster though slightly less accurate. A major disadvantage of MKL is the reliance on training data during testing which leads to slow inference and high memory usage.

Poria et al. [136] designed a real-time multimodal sentiment understanding model to harvest sentiments from web videos. They used feature and decision-level fusion methods to merge affective information extracted from multiple modalities. The audio, video and textual features are merged using decision-level and feature-level fusion. The text sentiment analysis model has been enriched by sentic computing-based features which significantly improved the performance of the model.

Rosas et al. [137] addressed a multimodal sentiment analysis framework on the Spanish videos. They prepared a dataset of 105 Spanish videos of 2-8 minutes length considering 21 male and 84 female speakers. Each video was segmented manually for 30 seconds that covered a single topic. They trained the model using a support vector classifier. The modality fusion of three different feature types leads to remarkable improvement rather than using the single modality. Due to the dataset constraint, the proposed approach detected the positive or negative polarity but could not identify the neutral polarity from the sentences.

Wöllmer et al. [138] carried out a multimodal sentiment analysis on online user-generated movie review videos. They collected 370 videos from YouTube and ExpoTV11 and manually annotated them into positive, negative and

neutral sentiment polarity. They used an automatic speech recognition (ASR) model and openSMILE software for text and audio feature extraction. Support vector machine and bidirectional long-short-term-memory (Bi-LSTM) classifiers were used to train linguistic and audiovisual features. Finally, a multimodal fusion technique was employed for sentiment prediction.

Majumder et al. [139] addressed the multimodal sentiment analysis framework using a context-aware hierarchical fusion scheme. A hierarchical fusion strategy first fuses the modalities two in two and only then fuses all three modalities. After extracting the text, audio and visual features, they transformed the context-aware utterance vectors into vectors of the same dimension. They compared and combined each bimodal combination of these abstract features using fully connected layers. Finally, they combined these bimodal vectors into a trimodal vector using fully-connected layers and used an RNN to pass contextual information between them.

Siddiquie et al. [140] explored a multimodal architecture to detect the sentiment of politically persuasive content on Youtube videos. They extracted the textual, audio and visual features in politically persuasive web videos to capture their affective semantics and the sentiment-related information in viewers' comments. For the video analysis, they used the ImageNet concepts and VSO concepts. Experimental results revealed that each feature modality could be used for politically persuasive content classification and the best performance could be obtained by fusing multiple modalities.

Pereira et al. [141] presented a multimodal sentiment prediction model on a news video dataset. They performed three different computational approaches: automatic emotion recognition from facial expressions, extraction of modulations in the participants' speeches and sentiment analysis from the closed caption associated with the videos of interest. A dataset containing 520 annotated news videos from American and Brazilian popular TV newscasts were used to experiment. Ellis et al. [142] also performed a multimodal sentiment prediction framework on broadcast news videos. They presented a 929 sentence-length videos in the domain of broadcast video news and performed annotation on the Amazon Mechanical Turk tool. Data was labeled into positive, negative and neutral sentiment polarity. Their findings revealed the importance of multimodal analysis for multimedia content in understanding its polarity and automatic sentiment analysis.

Huddar et al. [143] proposed an attention-based multimodal framework using a contextual fusion approach on IEMOCAP and CMU-MOSI datasets. They merged two-two modules simultaneously after extracting the features from the text, audio and video modalities. A bidirectional LSTM layer was employed to develop the classifier. Inter-modality and intra-modality dynamics are the most common challenges in multimodal sentiment analysis. Inter-modality dynamics is the interaction behavior change between language, i.e., visual and acoustic behaviors. The other is intra-modality to explore the sentiment/emotion efficiency of all the audio, video and acoustic modalities. This problem was addressed by Zadeh et al. [144] by developing a tensor fusion network. Unimodal, bimodal and trimodal interactions were explicitly merged in the inter-modality model. The acoustic, language and visual modalities were modeled through the intra-modality dynamics.

6.3.1 Multimodal Embedding. The multimodal semantics of words emphasizes embeddings with perceptual input assuming that human meaning representations are rooted in sensory perceptions. Khare et al. [145] extend the natural language understanding work using multimodal architectures that use textual, visual and audio information for machine learning tasks. Encoders from the transformer model were trained using multi-task training to extract the embeddings from their network. The author used automatic speech recognition and person identification as a part of their embedding generation framework. They fine-tuned and evaluated the embeddings on the CMU-MOSEI dataset for emotion recognition revealing that a cross-modal attention-based transformer architecture could improve the system's performance.

Similarly, Verő et al. [146] examined the impact of visual information on the semantics similarity and relatedness when the evaluation involves no direct visual input. They investigated a new embedding approach between visual and linguistic modalities based on the structured annotations of the Visual Genome. They also compared

unimodal and multimodal approaches including linguistic, structured and image-based representations. They revealed that the new embedding provides complementary information for text-based embeddings.

Mao et al.[147] focused on training and evaluating effective word embeddings with text and visual information. A large-scale dataset with 300 million sentences describing over 40 million images was introduced and crawled from Pinterest [1]. Comparing the various recurrent neural network (RNN) based multimodal (text and image) models demonstrated that incorporating the visual information into the word embeddings and a weight-sharing strategy is essential for multimodal embedding learning.

Niu et al. [148] reported a framework named cross-modality transfer learning (CMTL) which was related to distant domain transfer learning (DDTL) and negative transfer techniques. The conventional transfer learning approach assumes that the source and target domains are in the same modality. DDTL aims to make efficient transfers even when the domains are entirely different. Cross-modality transfer learning aims to make efficient transfers between two data modalities, i.e., from image to text. They aimed to improve image classification performance by transferring knowledge from text data. A summary of the selective multimodal sentiment analysis approaches is presented in Table 4.

Table 4. Summary of selective approaches of multimodal sentiment analysis

Author	Dataset	Approach/Findings	Performance
Morency et al. [133]	Youtube Dataset	Proposed a tri-modal sentiment analysis architecture by integrating the audio, video and text features for Spanish language	55.3%
Wollmer et al. [138]	Movie Reviews Video Dataset	A multimodal sentiment analysis framework by extracting text and audio features and applied SVM and Bi-LSTM classifiers	73.2%
Poria et al. [135]	video Dataset	Proposed a multimodal sentiment analysis model using multiple kernel learning approaches with an inner layer of deep CNN	88.6%
You et al. [126]	Getty, Twitter and VSO-VT Dataset	A semantic tree structure based on sentence parsing and integrates textual and visual information using an attention mechanism with LSTM for sentiment classification	90.2%
Pereira et al. [141]	News Video	Based on the fusion of audio, textual and visual clues extracted from different contents for multimodal sentiment analysis	84%
Majumder et al. [139]	MOSI, CMU-MOSI and IEMOCAP	Design a novel feature fusion strategy that proceeds hierarchically fusion approach for multimodal sentiment classification	80%
Huang et al. [127]	Getty Image, Twitter and Flickr Dataset	Design an image–text sentiment analysis framework for multimodal sentiment analysis.	86.9%
Huddar et al. [143]	IEMOCAP and CMU-MOSI Dataset	Presented an attention-based multimodal contextual fusion strategy for extracting the contextual information	86.5%

[1] https://www.pinterest.com/

6.4 Discussion

Remarkable progress has been made in joint visual-textual sentiment analysis in the multimodal domain using deep learning techniques. However, most of the existing studies on joint visual-textual sentiment analysis employ different fusion methods to integrate the textual and visual information ignoring the correlation between the textual and visual content. Although some researchers have begun to pay attention to this problem, there is still much room for further research. Research on multimodal sentiment analysis for audio-visual content in social media is particularly limited. This is because the research foundation of audio sentiment analysis is relatively weak and research on sentiment analysis for video data is in the early stage. Most existing studies focus primarily on self-timer web videos containing human faces.

7 RESOURCE OF SENTIMENT ANALYSIS RESEARCH

This section discusses sentiment analysis resources including lexicon and benchmark datasets.

7.1 Sentiment lexicon

Sentiment lexicons are commonly used resources for textual sentiment analysis. A lexicon is a list of units like words, phrases and word senses annotated by sentiment orientation. In a sentiment lexicon, each unit is treated as opinion information. Lexicon creation starts with an initial list of words also known as seed words which are extended using synonyms and antonyms of seed words. This process is repeated until the extension of the list is not stopped. Synonyms and antonyms words are taken from the WordNet dictionary. Lexicon can be broadly divided into two categories such as non-ontology-based and ontology-based methods. The non-ontology-based category includes lexica created using machine learning, lexicon-based and hybrid approaches. An ontology is an explicit specification of a conceptualization. It provides a formal representation of knowledge that enables reasoning. It is better than a taxonomy or relational database management system that captures semantic associations between concepts and relationships. Therefore, the sentiment analysis community is moving towards ontological-based approaches to represent a commonsense knowledge. General Inquirer (GI) is regarded as the earliest emotional thesaurus and affective analysis program in which emotional words are derived from Harvard IV-4 Dictionary and Lasswell's Dictionary. General Inquirer takes a pre-set list of positive and negative words and analyzes the polarity of documents based on the prevalence of words from each category. OpinionFinder is a sentiment lexicon including sources of opinion, direct subjective expressions, speech events and sentiment expressions. A brief description of the most widely used sentiment lexicons is shown in table 5.

Table 5. Summary of selective sentiment lexicon

Lexicon	Author	Size	Polarity Score
Harvard General Inquirer	Stone et al. [149],	11,790	positive, negative and neutral
WordNet-Affect (WNA)	Strapparava et al. [150]	1,903	positive, negative, neutral and ambiguous
SentiWordNet3.0	Baccianella et al. [30]	1,105	positivity, negativity and objectivity
MPQA	Wiebe et al. [151]	8,222	positive, negative and neural
Opinion lexicon	Liu et al. [152]	6,786	positive and negative
SenticNet	Poria et al. [153]	5,732	positive, negative, neutral and ambiguous

7.2 Dataset

The most challenging task for developing a machine learning and deep learning-based models is the presence of data from one particular domain. There are gold-standard benchmark datasets for sentiment analysis as discussed in table 6.

(1) **Yelp**: It is a five-class and two-class sentiment classification review dataset [2] namely Yelp-5 and Yelp-2. In the Yelp-5 dataset, the sentiment is classified into five fine-grained labels. Yelp-2 is annotated into positive and negative binary sentiment labels. The yelp-5 dataset has 650000 training and 50,000 testing samples and Yelp-2 has 560000 training and 38000 testing samples. Using the ALBERT model on the Yelp dataset, Alamoudi et al. [154] proposed an aspect-level sentiment classification framework using semantic similarity to leverage the powerful capacity of pre-trained language models and eliminate many complications associated with the supervised learning models. Food, service, ambiance and price are the aspects that have been categorized according to their sentiment context. An accuracy of 98.30% is reported.

(2) **IMDB**: IMDB [155] is a movie review dataset that is used for binary sentiment classification tasks [3]. IMDB has 25,000 reviews for training and 25,000 reviews for testing. Tang et al. [156] discussed document-level sentiment classification using CNN, LSTM and GRU deep learning models. They first used CNN or LSTM to generate sentence representations from word representations and then used GRU for encoding intrinsic relations between the sentences in the document. The experimental results showed an accuracy of 45.3% on the IMDB dataset.

(3) **Stanford Sentiment Treebank (SST)**: Socher et al. [157] developed the SST dataset which is an extension of the movie review (MR) dataset. The SST dataset has two versions. One is a five-class known as SST-1 and the other is a binary class known as SST-2. The fine-grained SST-1 includes 11,855 movie reviews partitioned into 8,544 training, 1,101 development and 2,210 testing samples. Again SST-2 consists of 6,920 training, 872 development and 1,821 testing samples. Chen et al. [158] reported a divide-and-conquer approach for sentence-level sentiment classification. Initially, target expressions from opinionated sentences are extracted using bi-LSTM with conditional random fields (biLSTM-CRF) and sentences are classified into a non-target group, one-target group and multi-target group. They conducted experiments on datasets the movie reviews, customer product reviews and stanford sentiment treebank (SST-1 and SST-2) with binary and fine-grained sentiment-labeled. An accuracy of 87.9% was achieved on the SST dataset using Bi-RNN and Bi-RNN with CRF approaches.

(4) **Amazon Review Dataset** [4]: Amazon review [159] is a product review dataset gathered from the Amazon website. The dataset is classified into binary and multi-class classifications depending on the labels. The binary classification consists of 3,600,000 training samples and 400,000 testing samples. There are 3,000,000 training and 650,000 testing datasets in the Amazon 5-class classification dataset. Haque et al. [160] applied supervised machine learning methods such as support vector machine (SVM), multinomial naïve Bayes (MNB), stochastic gradient descent (SGD), random forest (RF), decsision tree (DT) and logistic regression (LR) for sentiment classification on large-scale Amazon product reviews. They categorized the customer product reviews into positive and negative feedback and built a supervised learning model to polarize many reviews. The highest accuracy of 93.52% was achieved using 10-fold cross-validation on the electronics product.

(5) **SemEval 2007**: The dataset was developed by Strapparava et al. [161] for affective computing shared tasks in the SemEval 2007 workshop. The dataset comprises news headlines from major newspapers like the CNN, the New York Times, the BBC and the Google News search engine. Six annotators manually annotated the data into six different emotions: anger, fear, disgust, sadness, surprise and joy. Apart from binary labeling, they did fine-grained labeling for multiple emotions with different degrees on a scale of 0 to 100.

[2]https://www.kaggle.com/yelp-dataset/yelp-dataset
[3]https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
[4]https://www.kaggle.com/datafiniti/consumer-reviews-of-amazon-products

(6) **Getty Images Dataset**: Getty image dataset was presented by You et al. [162] which consists of 588,221 labeled image and text data. A tree LSTM with attention has the highest accuracy on the Getty images dataset.
(7) **CMU-MOSI**: CMU-MOSI [163] [5] is a multimodal corpus of sentiment intensity dataset which is a collection of 2199 opinion video clips. The opinion video is annotated with a sentiment ranging from -3 to 3. This dataset has been rigorously annotated with sentiment intensity, subjectivity, per-opinion, per-frame and per-millisecond visual features and audio features. Ghosal et al. [164] proposed a recurrent neural network-based multimodal attention framework that leverages contextual information for utterance-level sentiment prediction on the CMU-MOSI corpus. They demonstrated that the proposed method achieved accuracies of 82.31% of the CMU-MOSI dataset.
(8) **CMU-MOSEI:** The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [6] dataset is the most extensive sentiment analysis and emotion recognition dataset. More than 23,500 sentence utterance videos were collected from 1000 online YouTube speakers. All videos are gender-balanced. Each utterance is randomly selected from a variety of monologue videos and topics. Videos are transcribed and punctuated correctly. A gated mechanism for attention-based multimodal sentiment analysis was proposed by Kumar et al. [165] on CMU-MOSEI dataset. Their approach achieved an accuracy of 81.1% on the CMU-MOSEI multimodal corpus.
(9) **MOUD:** MOUD [166] is another multimodal opinion utterances dataset that consists of 80 videos in Spanish. Multiple segments in each video display either a positive, negative, or neutral sentiment. Wang et al. [167] proposed a Select-Additive Learning (SAL) procedure to improve the generalizability of trained neural networks for multimodal sentiment analysis. Their experiment using MOUD dataset yielded 57.4% accuracy.

Table 6. Summary of selective sentiment analysis dataset

Dataset	Domain	Model	F1-score and Accuracy
IMDB	Movie	LSTM+KNN (Zhou et al. [168]), **Joint RNN and CNN (Hasan et al. [169])**	69% F1-score and **93.2%** Accuracy
Yelp	Multi Review	LSTM+KNN (Zhou et al. [168]), **Bi-RNN + Attention (Zhang et al. [170])**	70% F1-score and **70.6%** Accuracy
SST	Movie	LSTM (Donnelly et al. [171]), **RNN + CNN (Hassan et al. [169])**	88.52% F1-score and **89.2%** Accuracy
Amazon	Product Review	Attention + Bi-LSTM (Yuan et al. [125]), **3 layers CNN + Bi-LSTM + Attention (Manushu et al. [172])**	87.6% F1-score and **87.7%** Accuracy
MOUD	Multimodal	LSTM + Hierarchical Fusion (Poria et al. [173]), **Bi-LSTM (Li et al. [174])**	68.11% F1-score and **71.1%** Accuracy
CMU-MOSI	Multimodal	Attention-based LSTM + Dynamic (Poria et al. [175]), **LSTM + Hierarchical Fusion (Poria et al. [173])**	75.1% F1-score and **81.3%** Accuracy

[5] https://www.kaggle.com/datasets/mathurinache/cmu-mosi
[6] http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/

8 PERFORMANCE EVALUATION PARAMETER

Most of the state-of-the-art sentiment analysis uses accuracy (Hassan et al. [169]), precision (Tay et al. [176]), recall (Zhou et al. [168]) and F1-score (Xiong et al. [177]). In general, accuracy is the most commonly used measure for assessing the performance of machine learning algorithms because it is adopted in almost all the studies in the sentiment analysis area. However, only using this indicator for performance evaluation is unfair due to the uneven samples in benchmark datasets. The number of positive and negative samples in the corpus may not be equal. To deal with this problem, researchers have adopted other measures such as precision, recall and F1-score to comprehensively evaluate. The precision is defined as the number of true positives over the number of true positives plus the number of false positives. The recall is defined as the number of true positives over the number of true positives plus the number of false negatives. The F1-Score measure provides a score that balances the concerns of precision and recall in one number. Tay et al. [176] has adopted accuracy evaluation measures for performance comparison in sentiment analysis. Niu et al. [178] also adopted accuracy and F1-score as system evaluation measures. You et al. [96] presented an image sentiment analysis framework and evaluated the prediction performance in terms of accuracy, F1-score, precision and recall.
Additionally, other parameters have been used for proper interpretation of results, such as mean square error (MSE) (Chaturvedi et al. [179]), ranking loss (Tang et al. [180]), macro-averaged mean absolute error (MAEM) (He et al. [181]) and RandIndex (Basha et al. [182]). To measure the divergence between predicted and actual sentiment labels, MSE can be used (Verma et al. [183]). The sentiment models are trained by minimizing the MSE (Jiang et al. [184]. Similarly, ranking loss measures the average distance between an actual sentiment value and the predicted sentiment value from sentiment classes with n number of test samples (Moghaddam et al. [185]). In order to handle imbalanced datasets, the macro-averaged mean absolute error is used as reported by Baccianella et al. [186]). Generally, RandIndex is used for clustering problems to determine whether two data clusters are similar as reported by (Zhao et al. [187]).

9 APPLICATIONS OF SENTIMENT ANALYSIS

Before the arrival of sentiment analysis, companies performed surveys or created focus groups to identify the product review which was much more expensive and slower. With the growth of opinions posted on social media, sentiment analysis has become a popular research domain. Researchers propose different techniques to predict the stock market by using sentiment analysis methods. Another important application for sentiment analysis is predicting the outcome of an election. Sentiment analysis helps the government or the political parties to get an idea about the election, their chance of winning, how much the public is satisfied, etc. It is considered the most helpful technique to predict the result of political decisions. Sentiment analysis can be helpful in the education field by improving teaching quality and student learning capacity. Other applications of sentiment analysis are box-office prediction, TV programs and news summarization, recommender systems, etc. A few domains have been explored yet and a lot more still needs to be explored. In future, we believe that much more exciting research on sentiment analysis will come forward with the emergence of various domains. A summary of sentiment analysis applications is provided in table 7.

10 CHALLENGES OF SENTIMENT ANALYSIS

Due to the diverse nature of human emotions, there are still lots of scope for developing unique systems or improving and enriching existing systems with effective modifications. Some possible issues related to the research of sentiment and opinion mining are listed in the following:

(1) **Sarcasm detection**: Sarcasm can be defined as saying or writing the opposite of what someone means or speaking in a way that makes someone feel stupid or angry. For example, when someone writes something positive but the meaning of the content is negative or vice versa that makes sentiment analysis more

Table 7. Summary of sentiment analysis applications

Domain	Application
Business	Brand reputation, E-commerce, Consumers voice, online advertising
Politics	Election outcome prediction, voting advise
Finance	Stock market prediction, Evaluation of commodities and share price
Public action	Intelligence transport system, Event monitoring
Healthcare	Medical field to assess the psychology and mental health
Education	Improving teaching quality, Understand student learning capacity
Multimedia analytics	News summarization, politically persuasive content identification, Recommender system

complex. Detecting sarcasm can be challenging due to the ambiguity and complexity of irony. Wen et al. [188] developed a sarcasm detection framework that outperformed the conventional deep learning models with an accuracy of 92.71%.

(2) **Handling negation**: Negation words such as not, neither, nor, etc., are essential for sentiment analysis because they can polarize a text. For example, "The food is tasty." is classified as positive polarity while "The food is not tasty." should be classified as negative polarity. Identifying implicit negation is one of the significant issues in negation handling tasks.

(3) **Low-resource language**: Most sentiment analysis models are data-driven; therefore preparing a standard labeled dataset is an essential task. The types of languages that suffer from linguistic resource scarcity are low-resource languages. In building dataset, data collected from various domains are often unstructured, noisy and wrongly spelled. Therefore, pre-processing steps are carried out to convert the data into machine consumable format which is time-consuming. Again, data annotation is a very time-consuming task and it varies from person to person. Apart from the dataset, sentiment lexicon is also required to perform sentiment analysis. But, it may create a bias in higher-level analysis or decision-making. Thus, the lack of resources can be overcome by constructing linguistic resources from scratch using semi-supervised, unsupervised and transfer learning methods [189].

(4) **Computational cost**: Increasing the training data size and complicating the model will exponentially increase the model's computational cost. Therefore, a high-end GPU is required to train the deep model with an enormous corpus. The neural and attention architecture are computationally costly compared to machine learning classifiers such as naïve Bayes and support vector machines.

(5) **Popularity Dynamic Prediction**: Popularity dynamic prediction is the task of forecasting the popularity achieved by images shared through social media over time. It is interesting and a recent challenge that concerns multimodal sentiment analysis related to social media content popularity. The popularity of social images which is usually estimated at a precise instant of the post life cycle could be affected by the period of the post (i.e., how old the post is). The level of engagement of an image posted on a social network is usually referred to as image popularity. The importance of image cues such as gradients, color, the set of objects present, deep learning features and the importance of various social cues such as the number of friends or the number of photos uploaded that lead to the high or low popularity of images. The challenge of predicting the popularity dynamics of social images has been introduced by Ortis et al. [11]. The correlation between social image popularity, visual features and social features extracted from images has been investigated by the author.

(6) **Virality:** Beyond the automatic understanding of objective properties of images like the presence of an object and its position in the scene, the computer vision community invested efforts in analyzing subjective

attributes of visual content. Virality is an example of such an attribute. In the overly-connected world, the automatic recognition of virality where the quality of an image or video can be rapidly and widely spread in social networks has crucial importance and recently awakened the interest of the computer vision community [190]. Social networking content virality is an important but esoteric phenomenon often studied in marketing, psychology and data mining.

(7) **Intra-modality Dynamics**: Another challenge in multimodal sentiment analysis is exploring the intra-modality dynamics of a specific modality. Intra-modality dynamics are particularly challenging for language analysis when performed on spoken language. An example of intra-modality dynamics is the utterance "This movie is sick," which can be ambiguous (either positive or negative) by itself, but if the speaker is also smiling at the same time, then it will be perceived as positive. On the other hand, the same utterance with a frown would be perceived negatively. A person speaking loudly, "This movie is sick," would still be ambiguous. Visual and acoustic modalities also contain their intra-modality dynamics which are expressed through both space and time [144].

(8) **Inter-modality dynamics**: The second challenge in multimodal sentiment analysis is exploring the inter-modality dynamics of a specific modality (unimodal interaction). The central challenge in multimodal sentiment analysis is to model the inter-modality dynamics: the interactions between language, visual and acoustic behaviors that change the perception of the expressed sentiment. Inter-modality dynamics are particularly challenging for language analysis when multimodal sentiment analysis is performed on spoken opinion. A spoken opinion such as "I think it was alright . . . Hmmm . . . let me think . . . yeah . . . no . . . ok yeah" rarely happens in written text. This volatile nature of spoken opinions complicate the proper language structure.

(9) **Emotion intensity detection:** An emotion may have different levels of intensity thereby detecting the intensity can be beneficial for emotion analysis. Considering "I am sad" or "I want to end my life; there is nothing left for me", the system will label both of these declarations as 'sad' emotions. However, the intensity of sadness in both statements is different.

(10) **Multiple emotion detection:** In most emotion detection, efforts mainly concentrate on the primary emotion in the text. Sentences expressing multiple emotions are labeled with the first detected emotion or discarded. For example, "I was sad this morning, but now I am happy," or "This makes me happy and sad at the same time," the system recognizes both emotions with a temporal and spatial dimension.

(11) **Emotion-cause detection**: Detecting the cause of emotion can increase the accuracy of detecting the correct emotion in a text or speech. For example, "I am so happy! It is raining!" the system recognizes the reason for his happiness is the rain.

(12) **Personality or mood detection**: By detecting the emotion from the text for a particular person, his personality or mood can be detected and analyzed. Personality or mood detection can be integrated into social networking platforms and other applications for personalized suggestions.

11 CONCLUSION AND FUTURE RESEARCH DIRECTION

We present the essentials of the state-of-the-art research works reported in sentiment analysis based on traditional machine learning and deep learning techniques. We extended the discussion focusing on the computational approaches for sentiment analysis considering different modalities including text, image, audio and different possible bi-modal and tri-modal combinations of the multimodal sentiment analysis. In addition, sentiment analysis resources and datasets are also elaborated. We give a brief overview of the applications and the evaluation measures of sentiment analysis. The sentiment analysis field is proven to be beneficial for politics, healthcare, education, product reviews, stock market prediction etc. With the rise in multi modal contents in the social media, it calls for a more complex multimodal sentiment analysis in order to develop various social media analytical

tools. In future, it is believed that more exciting research on sentiment analysis will emerge with the availability of various multimodal datasets on various domains. Exploiting the multimodal and multilingual features for sentiment analysis can further bolster cross-disciplinary research which can impact the technological advances ranging from business to politics and education to healthcare sector.

ACKNOWLEDGMENT

We acknowledge the CNLP (Centre for Natural Language Processing) at NIT Silchar for giving access to the lab.

REFERENCES

[1] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[2] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–33.

[3] Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*. 70–77.

[4] Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*. 519–528.

[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 79–86.

[6] Peter Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 417–424. DOI : http://dx.doi.org/10.3115/1073083.1073153

[7] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Twenty-Fourth international joint conference on artificial intelligence*.

[8] Cambria Erik, Poria Soujanya, Gelbukh Alexander, and Thelwall Mike. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32, 6 (2017), 74–80.

[9] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. 2020. Survey on visual sentiment analysis. *IET Image Processing* 14, 8 (2020), 1440–1456.

[10] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn Schuller, and Kurt Keutzer. 2018. Affective image content analysis: A comprehensive survey. (2018).

[11] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. 2019. An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges. *ICETE (1)* (2019), 296–306.

[12] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing* (2020).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Yinhan Liu, Myle Ott, Naman Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 1907. Roberta: A robustly optimized bert pretraining approach. arXiv 2019. *arXiv preprint arXiv:1907.11692* (1907).

[15] Zuhe Li, Yangyu Fan, Bin Jiang, Tao Lei, and Weihua Liu. 2019. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications* 78, 6 (2019), 6939–6967.

[16] Thoudam Doren Singh, Surmila Thokchom, Laiphrakpam Dolendro Singh, and Bunil Kumar Balabantaray. 2023. Recent advances on social media analytics and multimedia systems: issues and challenges. (2023).

[17] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. 2019. A survey of computational approaches and challenges in multimodal sentiment analysis. *Int. J. Comput. Sci. Eng* 7, 1 (2019), 876–883.

[18] Ganesh Chandrasekaran, Tu N Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 5 (2021), e1415.

[19] Ramandeep Kaur and Sandeep Kautish. 2022. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines* (2022), 1846–1870.

[20] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2022. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* (2022).

[21] Thoudam Doren Singh, Cristina España i Bonet, Sivaji Bandyopadhyay, and Josef van Genabith (Eds.). 2021. *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*. INCOMA Ltd., Online (Virtual Mode).

https://aclanthology.org/2021.mmtlrl-1.0
[22] Amitava Das and Sivaji Bandyopadhyay. 2010. Opinion-polarity identification in bengali. In *International Conference on Computer Processing of Oriental Languages*. 169–182.
[23] Kishorjit Nongmeikapam, Dilipkumar Khangembam, Wangkheimayum Hemkumar, Shinghajit Khuraijam, and Sivaji Bandyopadhyay. 2014. Verb based manipuri sentiment analysis. *Int J Nat Lang Comput* 3, 3 (2014), 113–118.
[24] Ringki Das and Thoudam Doren Singh. 2021. A Step Towards Sentiment Analysis of Assamese News Articles Using Lexical Features. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, Vol. 170. Springer, 15.
[25] Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation?. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*. 52–60.
[26] Kamil Kanclerz, Piotr Miłkowski, and Jan Kocoń. 2020. Cross-lingual deep neural transfer learning in sentiment analysis. *Procedia Computer Science* 176 (2020), 128–137.
[27] Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. 976–983.
[28] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. 34–35.
[29] Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th international conference on data engineering workshop*. IEEE, 507–512.
[30] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.. In *Lrec*, Vol. 10. 2200–2204.
[31] Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation* 8, 4 (2016), 757–771.
[32] Alexandra Balahur and Marco Turchi. 2013. Improving sentiment analysis in twitter using multilingual machine translated data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 49–55.
[33] Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. 2011. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Asia information retrieval symposium*. Springer, 238–249.
[34] Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 572–581.
[35] Ruifeng Xu, Jun Xu, and Xiaolong Wang. 2011. Instance level transfer learning for cross lingual opinion analysis. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. 182–188.
[36] Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1855–1862.
[37] Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. 2017. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review* 48, 4 (2017), 499–527.
[38] Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences* 512 (2020), 1078–1102.
[39] Denilson Alves Pereira. 2021. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review* 54, 2 (2021), 1087–1115.
[40] Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*. Association for Computing Machinery, New York, NY, USA, 168–177. DOI: http://dx.doi.org/10.1145/1014052.1014073
[41] Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband. 2018. Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications* 23, 1 (2018), 11.
[42] Wei Zhao, Ziyu Guan, Long Chen, Xiaofei He, Deng Cai, Beidou Wang, and Quan Wang. 2017. Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 30, 1 (2017), 185–197.
[43] Hua Xu, Fan Zhang, and Wei Wang. 2015. Implicit feature identification in Chinese reviews using explicit topic mining model. *Knowledge-Based Systems* 76 (2015), 166–175.
[44] Wanxiang Che, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu. 2015. Sentence compression for aspect-based sentiment analysis. *IEEE/ACM Transactions on audio, speech, and language processing* 23, 12 (2015), 2111–2124.
[45] Md Shad Akhtar, Tarun Garg, and Asif Ekbal. 2020. Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing* 398 (2020), 247–256.
[46] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.

[47] Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. 1367–1373.
[48] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[49] Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 conference on empirical methods in natural language processing*. 599–608.
[50] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.
[51] Pooja Pandey and Sharvari Govilkar. 2015. A framework for sentiment analysis in Hindi using HSWN. *International Journal of Computer Applications* 119, 19 (2015).
[52] Guang Qiu, Xiaofei He, Feng Zhang, Yuan Shi, Jiajun Bu, and Chun Chen. 2010. DASA: dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications* 37, 9 (2010), 6182–6191.
[53] Yao Lu, Xiangfei Kong, Xiaojun Quan, Wenyin Liu, and Yinlong Xu. 2010. Exploring the sentiment strength of user reviews. In *International Conference on Web-Age Information Management*. Springer, 471–482.
[54] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.
[55] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts *(ACL '04)*. Association for Computational Linguistics, USA, 271–es. DOI: http://dx.doi.org/10.3115/1218955.1218990
[56] Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence - Volume 2July 2006 Pages 1265–1270, pp. 1265–1270*. 1–6.
[57] Kamal Nigam and Matthew Hurst. 2004. Towards a robust metric of opinion. In *AAAI spring symposium on exploring attitude and affect in text*, Vol. 598603.
[58] Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *AMIA annual symposium proceedings*, Vol. 2005. American Medical Informatics Association, 570.
[59] Ziqiong Zhang, Qiang Ye, Zili Zhang, and Yijun Li. 2011. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications* 38, 6 (2011), 7674–7682.
[60] Thoudam Doren Singh, Telem Joyson Singh, Mirinso Shadang, and Surmila Thokchom. 2021. Review Comments of Manipuri Online Video: Good, Bad or Ugly. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, Vol. 170. Springer, 45.
[61] Loitongbam Sanayai Meetei, Thoudam Doren Singh, Samir Kumar Borgohain, and Sivaji Bandyopadhyay. 2021. Low resource language specific pre-processing and features for sentiment analysis task. *Language Resources and Evaluation* (2021), 1–23.
[62] Soujanya Poria, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems* 69 (2014), 45–63.
[63] Orestes Appel, Francisco Chiclana, Jenny Carter, and Hamido Fujita. 2016. A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level. *Know.-Based Syst.* 108, C (Sept. 2016), 110–124. DOI: http://dx.doi.org/10.1016/j.knosys.2016.05.040
[64] Alvaro Ortigosa, José M Martín, and Rosa M Carro. 2014. Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior* 31 (2014), 527–541.
[65] Lina Maria Rojas-Barahona. 2016. Deep learning for sentiment analysis. *Language and Linguistics Compass* 10, 12 (2016), 701–719.
[66] Peerapon Vateekul and Thanabhat Koomsubha. 2016. A study of sentiment analysis using deep learning techniques on Thai Twitter data. In *2016 13th international joint conference on computer science and software engineering (JCSSE)*. IEEE, 1–6.
[67] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18, 7 (2006), 1527–1554.
[68] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
[69] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. 2018. Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications* 114 (2018), 107–118.
[70] Hannah Kim and Young-Seob Jeong. 2019. Sentiment classification using convolutional neural networks. *Applied Sciences* 9, 11 (2019), 2347.
[71] Qiongxia Huang, Xianghan Zheng, Riqing Chen, and Zhenxin Dong. 2017. Deep Sentiment Representation Based on CNN and LSTM. In *International Conference on Green Informatics (ICGI),pp. 30-33*. 30–33.
[72] Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. 2019. A CNN-BiLSTM model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction* 1, 3 (2019), 832–847.
[73] K Shalini, Aravind Ravikurnar, Aravinda Reddy, KP Soman, et al. 2018. Sentiment analysis of indian languages using convolutional neural networks. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 1–4.

[74] Subhra Jyoti Baroi, Nivedita Singh, Ringki Das, and Thoudam Doren Singh. 2020. NITS-Hinglish-SentiMix at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text Using an Ensemble Model. (Dec. 2020), 1298–1303. https://aclanthology.org/2020.semeval-1.175

[75] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[76] Guozheng Rao, Weihang Huang, Zhiyong Feng, and Qiong Cong. 2018. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing* 308 (2018), 49–57.

[77] Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. 2019. A survey of sentiment analysis based on transfer learning. *IEEE Access* 7 (2019), 85401–85412.

[78] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* (2022), 1–50.

[79] Steve Yang, Jason Rosenfeld, and Jacques Makutonin. 2018. Financial aspect-based sentiment analysis using deep representations. *arXiv preprint arXiv:1808.07931* (2018).

[80] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 328–339.

[81] Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd nordic conference on computational linguistics*. 187–196.

[82] Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Improving BERT performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731* (2020).

[83] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. DOI: http://dx.doi.org/10.18653/v1/N18-1202

[84] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2022. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing* (2022).

[85] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric Xing. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. *Advances in neural information processing systems* 23 (2010).

[86] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. 2014. Object-Based Visual Sentiment Concept Analysis and Application. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*. Association for Computing Machinery, New York, NY, USA, 367–376. DOI: http://dx.doi.org/10.1145/2647868.2654935

[87] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. 2013. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. 1–8.

[88] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. 223–232.

[89] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*. 459–460.

[90] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[91] Donglin Cao, Rongrong Ji, Dazhen Lin, and Shaozi Li. 2016. Visual sentiment topic model based microblog image sentiment analysis. *Multimedia Tools and Applications* 75, 15 (2016), 8955–8968.

[92] Zuhe Li, Yangyu Fan, Weihua Liu, and Fengqin Wang. 2018. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multimedia Tools and Applications* 77, 1 (2018), 1115–1132.

[93] Yan-Ying Chen, Tao Chen, Winston H Hsu, Hong-Yuan Mark Liao, and Shih-Fu Chang. 2014. Predicting viewer affective comments based on image content in social media. In *proceedings of international conference on multimedia retrieval*. 233–240.

[94] Alessandro Ortis, Giovanni Maria Farinella, Giovanni Torrisi, and Sebastiano Battiato. 2020. Exploiting objective text description of images for visual sentiment analysis. *Multimedia Tools and Applications* (2020), 1–24.

[95] Zuhe Li, Yangyu Fan, and Weihua Liu. 2015. The effect of whitening transformation on pooling operations in convolutional autoencoders. *EURASIP Journal on Advances in Signal Processing* 2015, 1 (2015), 1–11.

[96] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI conference on artificial intelligence*.

[97] Stuti Jindal and Sanjay Singh. 2015. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In *2015 International Conference on Information Processing (ICIP)*. IEEE, 447–451.

[98] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[99] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing* 65 (2017), 15–22.
[100] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*. 675–678.
[101] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. 2010. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*. 715–718.
[102] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.
[103] Jingwen Wang, Jianlong Fu, Yong Xu, and Tao Mei. 2016. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks.. In *IJCAI*. 3484–3490.
[104] Zuhe Li, Qian Sun, Qingbing Guo, Huaiguang Wu, Lujuan Deng, Qiuwen Zhang, Jianwei Zhang, Huanlong Zhang, and Yu Chen. 2021. Visual sentiment analysis based on image caption and adjective–noun–pair description. *Soft Computing* (2021), 1–13.
[105] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*. 159–168.
[106] Hongyi Liu, Brendan Jou, Tao Chen, Mercan Topkara, Nikolaos Pappas, Miriam Redi, and Shih-Fu Chang. 2016. Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. 417–420.
[107] Unaiza Ahsan, Munmun De Choudhury, and Irfan Essa. 2017. Towards using visual attributes to infer image sentiment of social events. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1372–1379.
[108] Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
[109] Quanzeng You, Hailin Jin, and Jiebo Luo. 2017. Visual sentiment analysis by attending on local image regions. In *Thirty-First AAAI Conference on Artificial Intelligence*.
[110] Kaikai Song, Ting Yao, Qiang Ling, and Tao Mei. 2018. Boosting image sentiment analysis with visual attention. *Neurocomputing* 312 (2018), 218–228.
[111] Souraya Ezzat, Neamat El Gayar, and Moustafa M Ghanem. 2012. Sentiment analysis of call centre audio conversations using text classification. *International Journal of Computer Information Systems and Industrial Management Applications* 4, 1 (2012), 619–627.
[112] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2015. Automatic audio sentiment extraction using keyword spotting. In *Sixteenth Annual Conference of the International Speech Communication Association*.
[113] Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2017. Automatic sentiment detection in naturalistic audio. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 8 (2017), 1668–1679.
[114] Shahin Amiriparian, Nicholas Cummins, Sandra Ottl, Maurice Gerczuk, and Björn Schuller. 2017. Sentiment analysis using image-based deep spectrum features. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 26–29.
[115] Harika Abburi, Manish Shrivastava, and Suryakanth V Gangashetty. 2016. Improved multimodal sentiment detection using stressed regions of audio. In *2016 IEEE Region 10 Conference (TENCON)*. IEEE, 2834–2837.
[116] Sebastian Sager, Damian Borth, Benjamin Elizalde, Christian Schulze, Bhiksha Raj, Ian Lane, and Andreas Dengel. 2016. Audiosentibank: Large-scale semantic ontology of acoustic concepts for audio content analysis. *arXiv preprint arXiv:1607.03766* (2016).
[117] José Pereira, Jordi Luque, and Xavier Anguera. 2014. Sentiment retrieval on web reviews using spontaneous natural speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4583–4587.
[118] Min Wang, Donglin Cao, Lingxiao Li, Shaozi Li, and Rongrong Ji. 2014. Microblog Sentiment Analysis Based on Cross-Media Bag-of-Words Model. In *Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS '14)*. Association for Computing Machinery, New York, NY, USA, 76–80. DOI: http://dx.doi.org/10.1145/2632856.2632912
[119] Chao Chen, Fuhai Chen, Donglin Cao, and Rongrong Ji. 2015. A Cross-Media Sentiment Analytics Platform For Microblog *(MM '15)*. Association for Computing Machinery, New York, NY, USA, 767–769. DOI: http://dx.doi.org/10.1145/2733373.2807398
[120] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks *(MM '15)*. Association for Computing Machinery, New York, NY, USA, 1071–1074. DOI: http://dx.doi.org/10.1145/2733373.2806284
[121] Ziyuan Zhao, Huiying Zhu, Zehao Xue, Zhao Liu, Jing Tian, Matthew Chin Heng Chua, and Maofu Liu. 2019. An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management* 56, 6 (2019), 102097.
[122] Pengfei Li, Peixiang Zhong, Jiaheng Zhang, and Kezhi Mao. 2020. Convolutional Transformer with Sentiment-aware Attention for Sentiment Analysis. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–8. DOI: http://dx.doi.org/10.1109/IJCNN48605.2020.9206796
[123] Ringki Das and Thoudam Doren Singh. 2022. A multi-stage multimodal framework for sentiment analysis of Assamese in low resource setting. *Expert Systems with Applications* (2022), 117575.

[124] Ringki Das and Thoudam Doren Singh. 2023. Image-Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (feb 2023). DOI: http://dx.doi.org/10.1145/3584861 Just Accepted.
[125] Zhigang Yuan, Sixing Wu, Fangzhao Wu, Junxin Liu, and Yongfeng Huang. 2018. Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems* 155 (2018), 1–10.
[126] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. 2016. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *proceedings of the 24th ACM international conference on multimedia*. 1008–1017.
[127] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems* 167 (2019), 26–37.
[128] Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. IJCAI.
[129] Sumit K Yadav, Mayank Bhushan, and Swati Gupta. 2015. Multimodal sentiment analysis: Sentiment analysis using audiovisual format. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 1415–1419.
[130] Eric Chu and Deb Roy. 2017. Audio-visual sentiment analysis for learning emotional arcs in movies. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 829–834.
[131] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting emotions in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.
[132] Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain. 2021. A novel context-aware multimodal framework for persian sentiment analysis. *Neurocomputing* 457 (2021), 377–388.
[133] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. 169–176.
[134] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. 2017. Benchmarking multimodal sentiment analysis. In *International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, 166–179.
[135] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2539–2544.
[136] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.
[137] Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems* 28, 3 (2013), 38–45.
[138] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28, 3 (2013), 46–53.
[139] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems* 161 (2018), 124–133.
[140] Behjat Siddiquie, Dave Chisholm, and Ajay Divakaran. 2015. Exploiting Multimodal Affect and Semantics to Identify Politically Persuasive Web Videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 203–210. DOI: http://dx.doi.org/10.1145/2818346.2820732
[141] Moisés Henrique Ramos Pereira, Flávio Luis Cardeal Pádua, Adriano César Machado Pereira, Fabrício Benevenuto, and Daniel Hasan Dalip. 2016. Fusing audio, textual, and visual features for sentiment analysis of news videos. In *Tenth International AAAI Conference on Web and Social Media*.
[142] Joseph G. Ellis, Brendan Jou, and Shih-Fu Chang. 2014. Why We Watch the News: A Dataset for Exploring Sentiment in Broadcast Video News. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 104–111. DOI: http://dx.doi.org/10.1145/2663204.2663237
[143] Mahesh G Huddar, Sanjeev S Sannakki, and Vijay S Rajpurohit. 2021. Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. *Multimedia Tools and Applications* 80, 9 (2021), 13059–13076.
[144] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).
[145] Aparna Khare, Srinivas Parthasarathy, and Shiva Sundaram. 2020. Multi-modal embeddings using multi-task learning for emotion recognition. *arXiv preprint arXiv:2009.05019* (2020).
[146] Anita L Verő and Ann Copestake. 2021. Efficient Multi-Modal Embeddings from Structured Data. *arXiv preprint arXiv:2110.02577* (2021).
[147] Junhua Mao, Jiajing Xu, Kevin Jing, and Alan L Yuille. 2016. Training and evaluating multimodal word embeddings with large-scale web annotated images. *Advances in neural information processing systems* 29 (2016).
[148] Shuteng Niu, Yushan Jiang, Bowen Chen, Jian Wang, Yongxin Liu, and Houbing Song. 2021. Cross-modality transfer learning for image-text information management. *ACM Transactions on Management Information System (TMIS)* 13, 1 (2021), 1–14.
[149] Philip J. Stone and Earl B. Hunt. 1963. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference (AFIPS '63 (Spring))*. Association for Computing Machinery, New

York, NY, USA, 241–256. DOI : http://dx.doi.org/10.1145/1461551.1461583
[150] Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf
[151] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39, 2 (2005), 165–210.
[152] Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*. 342–351.
[153] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems* 28, 2 (2013), 31–38.
[154] Eman Saeed Alamoudi and Norah Saleh Alghamdi. 2021. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems* 30, 2-3 (2021), 259–281.
[155] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.
[156] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1422–1432.
[157] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170
[158] Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications* 72 (2017), 221–230.
[159] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*. 440–447.
[160] Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. 2018. Sentiment analysis on large scale Amazon product reviews. In *2018 IEEE international conference on innovative research and development (ICIRD)*. IEEE, 1–6.
[161] Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07)*. Association for Computational Linguistics, USA, 70–74.
[162] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM international conference on Web search and data mining*. 13–22.
[163] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
[164] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*. 3454–3466.
[165] Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4477–4481.
[166] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 973–982.
[167] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 949–954.
[168] Ke Zhou, Jiangfeng Zeng, Yu Liu, and Fuhao Zou. 2018. Deep sentiment hashing for text retrieval in social CIoT. *Future Generation Computer Systems* 86 (2018), 362–371.
[169] Abdalraouf Hassan and Ausif Mahmood. 2018. Convolutional recurrent deep learning model for sentence classification. *Ieee Access* 6 (2018), 13949–13957.
[170] Jia-Dong Zhang and Chi-Yin Chow. 2019. MOCA: multi-objective, collaborative, and attentive sentiment analysis. *IEEE Access* 7 (2019), 10927–10936.
[171] Jonathan Donnelly and Adam Roegiest. 2019. On interpretability and feature representations: an analysis of the sentiment neuron. In *European Conference on Information Retrieval*. Springer, 795–802.
[172] Tu Manshu and Wang Bing. 2019. Adding prior knowledge in hierarchical attention neural network for cross domain sentiment classification. *IEEE Access* 7 (2019), 32578–32588.
[173] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 873–883.

[174] Haoran Li and Hua Xu. 2019. Video-based sentiment analysis with hvnLBP-TOP feature and bi-LSTM. In *proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9963–9964.
[175] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1033–1038.
[176] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2017. Dyadic memory networks for aspect-based sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 107–116.
[177] Shufeng Xiong, Hailian Lv, Weiting Zhao, and Donghong Ji. 2018. Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing* 275 (2018), 2459–2466.
[178] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*. Springer, 15–27.
[179] Iti Chaturvedi, Ranjan Satapathy, Sandro Cavallari, and Erik Cambria. 2019. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters* 125 (2019), 264–270.
[180] Duyu Tang, Bing Qin, and Ting Liu. 2015. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5, 6 (2015), 292–303.
[181] Yunchao He, Liang-Chih Yu, Chin-Sheng Yang, K Robert Lai, and Weiyi Liu. 2016. YZU-NLP team at semeval-2016 task 4: Ordinal sentiment classification using a recurrent convolutional network. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 251–255.
[182] Syed Muzamil Basha and Dharmendra Singh Rajput. 2019. A roadmap towards implementing parallel aspect level sentiment analysis. *Multimedia Tools and Applications* 78, 20 (2019), 29463–29492.
[183] Sharad Verma, Mayank Saini, and Aditi Sharan. 2017. Deep sequential model for review rating prediction. In *2017 Tenth International Conference on Contemporary Computing (IC3)*. IEEE, 1–6.
[184] Mengxiao Jiang, Jianxiang Wang, Man Lan, and Yuanbin Wu. 2017. An effective gated and attention-based neural network model for fine-grained financial target-dependent sentiment analysis. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 42–54.
[185] Samaneh Moghaddam and Martin Ester. 2010. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1825–1828.
[186] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Multi-facet rating of product reviews. In *European conference on information retrieval*. Springer, 461–472.
[187] Li Zhao, Minlie Huang, Haiqiang Chen, Junjun Cheng, and Xiaoyan Zhu. 2014. Clustering aspect-related phrases by leveraging sentiment distribution consistency. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1614–1623.
[188] Zhiyuan Wen, Lin Gui, Qianlong Wang, Mingyue Guo, Xiaoqi Yu, Jiachen Du, and Ruifeng Xu. 2022. Sememe knowledge and auxiliary information enhanced approach for sarcasm detection. *Information Processing & Management* 59, 3 (2022), 102883.
[189] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* 226 (2021), 107134.
[190] Xavier Alameda-Pineda, Andrea Pilzer, Dan Xu, Nicu Sebe, and Elisa Ricci. 2017. Viraliency: Pooling local virality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6080–6088.