A Comprehensive Review of Visual-Textual Sentiment Analysis from Social Media Networks

ISRAA KHALAF SALMAN AL-TAMEEMI, Mohammad-Reza Feizi-Derakhshi
Computerized Intelligence Systems Laboratory
Department of Computer Engineering
University of Tabriz
Tabriz, Iran
{asra_salman, mfeizi}@tabrizu.ac.ir

Saeed Pashazadeh, Mohammad Asadpour
Department of Computer Engineering
University of Tabriz
Tabriz, Iran
{pashazadeh, m_asadpour}@tabrizu.ac.ir

Corresponding author: Mohammad-Reza Feizi-Derakhshi (e-mail: mfeizi@tabrizu.ac.ir).

Abstract

Social media networks have become a significant aspect of people's lives, serving as a platform for their ideas, opinions and emotions. Consequently, automated sentiment analysis (SA) is critical for recognising people's feelings in ways that other information sources cannot. The analysis of these feelings revealed various applications, including brand evaluations, YouTube film reviews and healthcare applications. As social media continues to develop, people post a massive amount of information in different forms, including text, photos, audio and video. Thus, traditional SA algorithms have become limited, as they do not consider the expressiveness of other modalities. By including such characteristics from various material sources, these multimodal data streams provide new opportunities for optimising the expected results beyond text-based SA. Our study focuses on the forefront field of multimodal SA, which examines visual and textual data posted on social media networks. Many people are more likely to utilise this information to express themselves on these platforms. To serve as a resource for academics in this rapidly growing field, we introduce a comprehensive overview of textual and visual SA, including data pre-processing, feature extraction techniques, sentiment benchmark datasets, and the efficacy of multiple classification methodologies suited to each field. We also provide a brief introduction of the most frequently utilised data fusion strategies and a summary of existing research on visual–textual SA. Finally, we highlight the most significant challenges and investigate several important sentiment applications.

Keywords Deep learning · Machine learning · Multimodal fusion · Sentiment analysis · Visual-textual sentiment classification

1 Introduction

Today's affordable and comprehensive Web provides plenty of valuable social data to enhance decision-making. Sentiment analysis (SA) has become increasingly important for tracking people's emotions and opinions by evaluating unorganised, high-dimensional, multimodal and noisy social data. SA aids in gathering information about a given subject or topic in different fields, such as business, marketing, and intelligence services. It is an area of natural language processing (NLP) that concentrates on analysing massive amounts of user-generated online content produced daily by users on social platforms, such as Flickr, Instagram, Twitter and Facebook. These networks contain a wide range of social content, enabling for the detection of sentiment reflected by textual and visual data. Nowadays, if one wishes to purchase a consumer product, one does not have to rely solely on the guidance of friends and family because many user reviews and comments about the product are available in public Web forums. For example, we may consult Amazon reviews to determine which product best suits our needs. The significance of reviews arises from their ability to correctly represent previous consumer experiences. However, identifying and monitoring online opinion sites and collecting the information contained inside seems to be a difficult task. A typical human reader has difficulty finding relevant websites, extracting their thoughts and summarising them; hence, automated SA tools are necessary. In addition to real-world applications, several application-oriented research articles on topics such as stock market forecasting [1], political election prediction [2] and even healthcare [3] have been published. In X. Li et al. [1] work the financial news articles were used to measure their influences on stock price return. A sentiment analysis strategy for forecasting the result of presidential elections in a Twitter nation was reported by V. Kagan et al. [2]. Social media data was utilised by S. Yadav et al. [3] to build a patient-assisted system via medical SA. As a result, significant work has been conducted on user SA using textual data for numerous applications, which is commonly referred to as the process of automatically evaluating whether a part of text is positive, negative, or neutral.

In addition to textual data, images play an important role in expressing users' emotions on various online social media platforms. With the emergence of social networks, images have become a simple and effective method of conveying information amongst online users. Similar to textual data, images transmit various levels of emotion to their viewers. Nevertheless, extracting and interpreting an image's sentiment remains a challenge. As a result, visual sentiment analysis (VSA) has become important for investigating online multimedia content. It involves the ability to recognise an object, action, scene and emotional content from an image and classify them into different sentiment polarities, which can be performed by portraying the image in terms of colour and utilising gist features to categorise images based on different emotions. Moreover, deep learning (DL) architectures have seen tremendous success in the computer vision field [4]. Convolutional neural networks (CNNs) have demonstrated robust and accurate feature learning capabilities compared with DL-based networks. CNN models are more similar to human performance in visual recognition; they have been used successfully in tasks such as feature learning [5], sentence classification [6] and image classification [7]. Given the rapid growth and popularity of social media, scholars have been able to broaden the scope of SA to include other interesting applications. For example, the vast majority of VSA study focuses on face close-up images, in which facial expressions are used as visual signals to infer sentiments and forecast emotions [8].

As previously mentioned, most of the prior research on textual SA has been widely investigated [61–122], revealing various methodologies for mining opinion, including machine learning (ML), lexicon-based and hybrid techniques using Twitter text data. Moreover, only a few relevant studies are available in the literature on VSA [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. However, most of the published research has focused on analysing data from a single modality, neglecting the complementary information included in visual and textual contents. In addition, human expressions are quite complicated because words, visuals and their combinations can be used to represent a wide range of emotions and frequently require context to be comprehended properly. Thus, understanding the inner relationship between image and text is important because this combination can change or improve the meanings and thus the sentiment.

Multimodal SA has recently received considerable attention from academics to cope with diverse social media trends, given that people are increasingly expressing themselves and sharing their experiences through multiple types of social media. The basic goal of multimodal SA is to capture users' feelings by combining expressive data from many modalities. It also focuses on building semantic linkages between user-provided contents, such as text, images, videos and other types of media. The current state of research in multimodal SA is divided into three primary categories:

- Social media image and tagged content analysis [24].
- Audio and visual data analysis [25].
- Human–human and human–machine interaction analysis [26].

Our review article will focus on the most significant field of multimodal SA, which works on visual and textual information posted on social media platforms. Many people are more inclined to use these data to express themselves on these platforms. The visual and textual characteristics will be integrated to establish the general attitudes conveyed in the postings, which will then be categorised as positive, negative or neutral.

1.1 SCOPE OF THIS SURVEY

Most previous surveys have ignored the majority of visual–textual SA techniques, focusing instead on supervised ML and DL techniques. Although the present work discusses these approaches, it differs from earlier research by providing a more comprehensive examination of SA because it covers many aspects of the discipline, such as challenges, applications, sentimental datasets, feature extraction techniques, data fusion approaches and classification strategies. Scholars and novices will benefit greatly from this study because they may obtain a wealth of knowledge on this topic in a single paper. Our study varies from previous studies by detailing the benefits and drawbacks of the most commonly used fusion approaches, which may help researchers in selecting the best answer to their problems. The following are some of the survey's most important contributions:

- It reviews current works to offer researchers with a complete understanding of the methodologies and resources available for visual and textual SA.
- It provides a comprehensive overview of visual and textual SA, including data pre-processing, feature extraction techniques and sentiment benchmark datasets.
- This study categorises and summarises the most common SA methodologies, namely, ML, lexicon-based, hybrid and DL methods.
- It provides a brief introduction of the most widely used data fusion strategies and summarises the existing research on visual–textual SA by referencing previously published works.
- It summarises the applications and challenges associated with SA.

The remainder of this paper is organised as follows: In Section II, sentiment definitions and terms are introduced. Section III presents the basic architecture of the visual–textual SA and thoroughly discusses the complete process, including critical steps, such as data pre-processing, feature extraction techniques and the most important SA classification algorithms for textual and visual components along with the most important measures used in this area. Section IV presents SA challenges, and Section V discusses the significant sentiment applications. Section VI concludes this paper.

1.2 MOTIVATION

The number of articles in SA and opinion mining has increased dramatically, making this topic one of the most researched topics. SA can assess, identify and classify people's emotions, attitudes, feelings and opinions conveyed in text. With the explosive expansion of social media networks (e.g. Twitter, Facebook and Instagram), multimedia information like text, images, audio and videos is crucial in conveying people's thoughts. Multimodal SA from social media is becoming a topic in computer vision, pattern recognition and other areas of artificial intelligence. However, most current SA research focuses on a single modality, making it difficult to discern people's actual sentiment. The accuracy, reliability and robustness of these single-mode systems are limited. Thus, the usage of several modalities should be investigated to improve SA systems.

This study aims to investigate practical and empirical aspects for blog of data that is accessible through social media networks (e.g. Twitter, Instagram, Facebook and other websites) that provide such multimedia data. The topic of SA has become a topic of interest due to the growing expansion of information on social media. Several aspects of multimedia business research are investigated, all of which benefit the business community as a whole. Another aim is to develop a mathematical framework for multimodal SA, which includes numerous inputs to improve SA's overall performance. Once implemented, the model will assist social media researchers in determining numerous sentiment-related factors. Our study focuses on providing a comprehensive description of the datasets, feature extraction methods, fusion approaches, classification methods and difficulties encountered while performing multimodal SA.

2 SENTIMENT DEFINITIONS AND TERMS

In NLP, the terms 'affect', 'feeling', 'emotion', 'sentiment', and 'opinion' are often interchanged. 'Emotion' has several descriptions and related concepts, many of which rely on Scherer's emotion theory [27]. According to Scherer, emotions are fleeting phenomena that require activation. They are composed of cognitive appraisals, physical responses, action inclinations (e.g. fight or flight), facial and vocal expressions and subjective experiences. The mood is a long-lasting term, broad emotional condition that occurs for no apparent reason and remains for hours or days. Feelings are emotional experiences that are unique to each individual.

Deonna and Teroni [28] defined affective phenomena philosophically; they defined sentiment as a deeply held belief that may take many forms, such as 'the passion you may have for your hamster, your allegiance to your nation, your contempt for the financial system and your extreme liking for the latest technological item'. Sentiment manifests only when the sentiment holder is met with an entity or object. According to Deonna and Teroni [28], emotional interactions with an object or entity may be utilised to identify or monitor sentiment. For example, if A Likes B, then A believes that B is in a good mood.

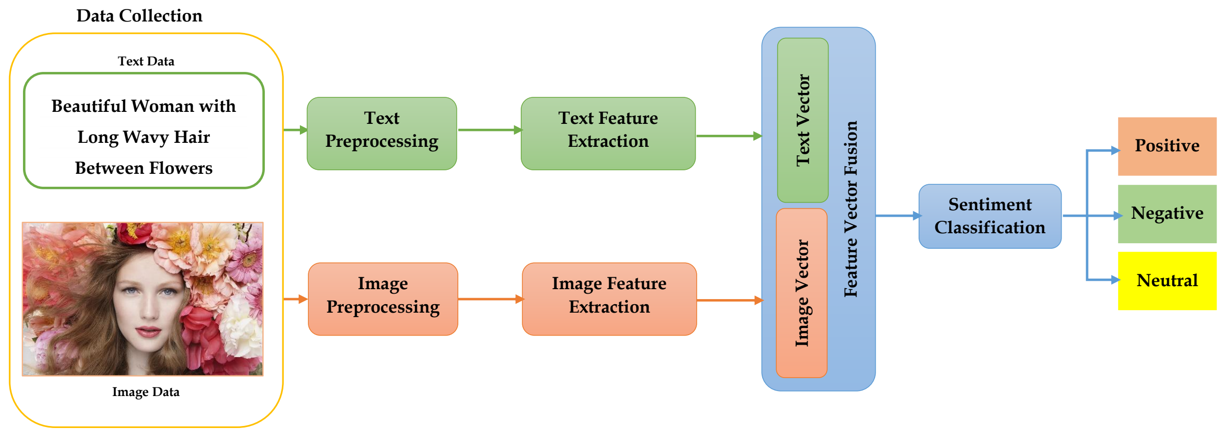

Figure 1: Basic architecture for visual–textual sentiment analysis.

Munezero et al. [29] comprehensively discussed emotion recognition and SA in text to distinguish between emotion and sentiment based on their duration. Emotions fade quickly, whereas sentiments stay for a long time. The authors described opinions as debatable and ambiguous assessments that do not have to be emotionally charged. SA is often associated with opinion mining, and it evaluates highly emotional attitudes and feelings [30]. In sum, different from emotions, sentiments and opinions may not always appear in behaviours or expressions. Emotions include a person, subjective feelings, physical changes and a target; whereas sentiments have a sentiment holder, polarity (positive or negative) and an object. Previously, there existed an interest to understand others' ideas and impressions. However, more organisations have shown interest in comprehending consumer attitudes to better satisfy their needs and enhance client acquisition and retention [31].

3 VISUAL-TEXTUAL SA FRAMEWORK

The basic architecture for visual–textual SA is shown in Figure 1. It comprises different modules, starting with the data collection module, which is the first phase in visual–textual SA. The data are in the form of texts, images from a dataset or social media posts. During the pre-processing phase, visual–textual SA prepares the data received from the data collection module for further processing. Each modality has unique features that must be treated separately, features are extracted from each modality independently and represented as vectors throughout the feature extraction and vector representation phases. These feature vectors are then combined to generate a single feature representation for recognising the emotion of the user's content. The combined feature vector is used in the classification phase to classify the data collected and processed into positive, negative or neutral. The last module of this design visualises and shows user attitudes based on the previous study. We will provide a brief review of each of these phases in the following sections.

3.1 TEXTUAL SENTIMENT ANALYSIS

SA, through social media content, aims to capture the sentiment information from social media posts. SA for social networks is more difficult than extracting sentiment from webpages (e.g. blogs and forums) due to the unique features of social media data. The brief phrasing and unstructured style of social media communication pose various issues. This text analysis allows businesses to determine client needs and political parties to conduct opinion surveys. Many applications rely on textual sentiment research, such as stock market forecasting, box office prediction and election forecasting. Three conventional textual SA levels are used: document, sentence and entity levels. Social media comments are usually no more than two lines long due to space constraints. In this scenario, document and sentence levels are the same. Therefore, two levels of textual SA need to be defined for social media, including message/sentence and entity levels [32]. This section outlines the various pre-processing operations, feature extraction techniques and computational approaches related to textual SA.

3.1.1 TEXT PRE-PROCESSING

Information obtained from different resources, particularly social media, is frequently unstructured. Raw data may be noisy and include spelling and grammatical errors [33]. Thus, texts must be cleaned up before analysing them. Given that many words are meaningless and add nothing to the text (e.g. stop words, prepositions, punctuation and special characters), pre-processing seeks to improve the analysis and minimise the dimensionality of the input data. Several typical tasks are included during the entire procedure as follows.

- Lowercasing. Changing all of the texts' letters to the same case. For example, the word Ball will be changed into ball.
- Substituting for negative words. Tweets include various negational concepts. The negation process converts won't and can't to will not and cannot, respectively.
- Eliminating unnecessary information [including punctuation, hashtags (#), special characters ($, &, %, ...), additional spaces, stop words (e.g. 'the', 'is', 'at', etc.), URL references, @username, numbers and non-ASCII-English characters to preserve the scope of the information that is unique to the English language]. Such information does not expect the expression of users' emotions.
- Translating emoticons. Nowadays, users utilise emoticons to express their thoughts, feelings and emotions. As a result, translating all emoticons to their corresponding words will produce better results.
- Changing words with repeated characters to their English origins. Individuals often employ words with repeated letters (e.g. 'coooool') to convey their feelings.
- Expanding acronyms to their original words using an acronym dictionary, Abbreviations and slang are poorly constructed words that are frequently used in tweets. They must be restored to their original form.
- Tokenisation. This stage deconstructs text into small units called tokens (e.g. documents into sentences, sentences into words).
- Part-of-speech (PoS) tagging. This stage detects the many structural parts in a text, including verbs, nouns, adjectives and adverbs.
- Lemmatisation. It is the process of reducing a particular word to its simplest form, identical to stemming, but it retains word-related information like PoS tags.

Several researchers have taken advantage of the pre-processing techniques. Zainuddin et al. [34] used text pre-processing approaches, including stemming and stop word removal, to analyse attitudes in texts. Sharma et al. [35] used PoS tagging to perform text-based SA on an online movie review dataset. Normalisation, URL removal and acronym expansion have also been considered. Sailunaz et al. [36] used these approaches to demonstrate how text preparation improves Twitter SA accuracy. Pradha et al. [37] performed text-based SA on a Twitter dataset using lemmatisation and stemming.

3.1.2 TEXT FEATURE EXTRACTION

Feature extraction, also known as feature engineering, is critical in the SA process since it affects sentiment classification performance directly [38, 39]. The goal of this assignment is to capture valuable information (e.g. phrases that represent emotions) from the text that explains important features of the text. The three types of feature extraction methods are as follows.

A set of Hand-crafted Features which are introduced by S. M. Mohammad et al. [40]. These features comprise the following:

- All Caps. The number of words that have all letters capitalised.
- Emoticons. A simple method of describing characteristics that consider the presence of positive (or negative) emoticons and whether the final unit of segmentation is an emoticon.
- Elongated Units. The number of elongated words (words that repeat one character more than twice; for example, 'gooood').
- Sentiment Lexicon. Various sentiment lexicons may be used to capture textual features by calculating the number of sentiment words, the score of the final sentiment word, the overall sentiment score and the maximum sentiment score for each lexicon.
- Cluster. The total number of words from each of the 1,000 clusters in the Twitter NLP tool [41].

- Ngrams (terms of presence and frequency). It is the most basic kind of feature encoding and is commonly used in information retrieval and SA. Features are defined as a single word or a list of n consecutive words that may be combined to make a unigram, bigram or trigram, and the frequency counts of those words. The term ‘presence’ assigns a binary value to the words (zero if the word is present, and one otherwise).
- PoS Tags. It represents the labels that define the function of word in a language. Words may be divided into numerous components of speech (e.g. noun, verb, article, adjective, preposition, pronoun, adverb, conjunction and interjection). For instance, ‘The phone is beautiful’ will be attached with a Stanford PoS tagger presented by K. Toutanova et al. [42]: The (determiner DT), phone (noun NN), is (verb VBZ), beautiful (adjective JJ). Adjectives are used as important opinion indicators in some SA methods [43].
- Opinion Words and Phrases. Opinion words are frequently used to convey positive or negative feelings (e.g. nice and amazing represent positive feelings; horrible and dreadful represent negative feelings).
- Negations. Negation words (known as opinion or valence shifters [44, 45]) are terms that can modify and reverse the emotional polarity. The most frequent negative words are ‘not’, ‘never’, ‘none’, ‘nobody’, ‘nowhere’, ‘neither’ and ‘cannot’. However, during the pre-processing stage, these terms are often regarded stop-words and are excluded from the text. Negation words should be treated with caution due to their importance because not all negated words result in negation.

Statistical Feature Representation. Here, input text is converted into a feature vector of fixed length that can be used in classification procedures. This method analyses the weights of words in a text using predefined keywords and generates a digital vector representing the text’s feature vector. Some significant text representation approaches are as follows.

- Bag-of-Words (BoW) model. It is one of the most basic and widely used strategy for translating text to numerical form (vector) [38]. Nevertheless, it loses the textual syntactic information, because it ignores word order, sentence structure or grammatical construction, focusing instead on word occurrence. The BoW model creates a vocabulary of all unique words in the document, and then it encodes any sentence as a fixed-length vector with the length of the vocabulary of known words, where the value of each position in the vector indicates a count or frequency of each word in the training set. A simple and efficient extension of the BoW is the term frequency–inverse document frequency (TF–IDF).
- TF–IDF. It is a statistical measure of the importance of a word to a document within a corpus of texts. This is achieved by multiplying two metrics: the frequency with which a word occurs in a document and its inverse document frequency over a group of documents. The TF–IDF value grows according to the number of times a word occurs in the document and is offset by the number of corpus documents that include the term, which helps to account for the fact that particular words occur more frequently than others. The TF–IDF is often computed using the scikit-learn library’s vectorizer class.

Distributed Word Representation. In contrast to the BoW model, the distributed representation (also referred to as word embedding) of a qualitative concept (e.g. word, paragraph or document) distributes the concept information along a vector. As a result, each point in the vector might indicate a non-zero value for a particular idea. Typically, distributed representations are used in conjunction with DL models. A quick overview of the most commonly used distributed representation approaches is presented as follows.

- Word2vec. It is a popular method for obtaining distributed representations of words; it was developed by Mikolov et al. [46] using shallow neural networks (NNs). Its architecture comprises the continuous BOW (CBOW) model, which estimates the current word based on the surrounding context words, and the Skip-Gram (SG) model, which uses the current word to estimate the context words.
- Global Vectors (GloVe). It is an unsupervised learning approach proposed by Pennington et al. [47] to create word embeddings from a corpus by collecting global word–word co-occurrence matrices to reveal important linear substructures of the word vector space. Due to its parallel implementation, the GloVe model may be trained efficiently on bigger datasets [48].
- FastText. It is a novel technique based on Bojanowski’s SG model [49] that treats each word as a bag of character n-grams. Each letter n-gram has a vector representation, with words encoded as the sum of these representations. It is rapid, which allows the models to be quickly trained on big corpora and word representations to be calculated for terms that are not in the training data.

Several researchers have used different types of feature extraction methods to retrieve the most important textual information. Jabreel and Moreno [50] proposed DL model that solved the emotion classification problem in twitter data by exploiting word2vec method for extracting the tweet features. To extract text features, Kaibi et al. [51] utilised feature extraction models, namely, word2vec, a prominent word embedding model based on NN architectures; GloVe;

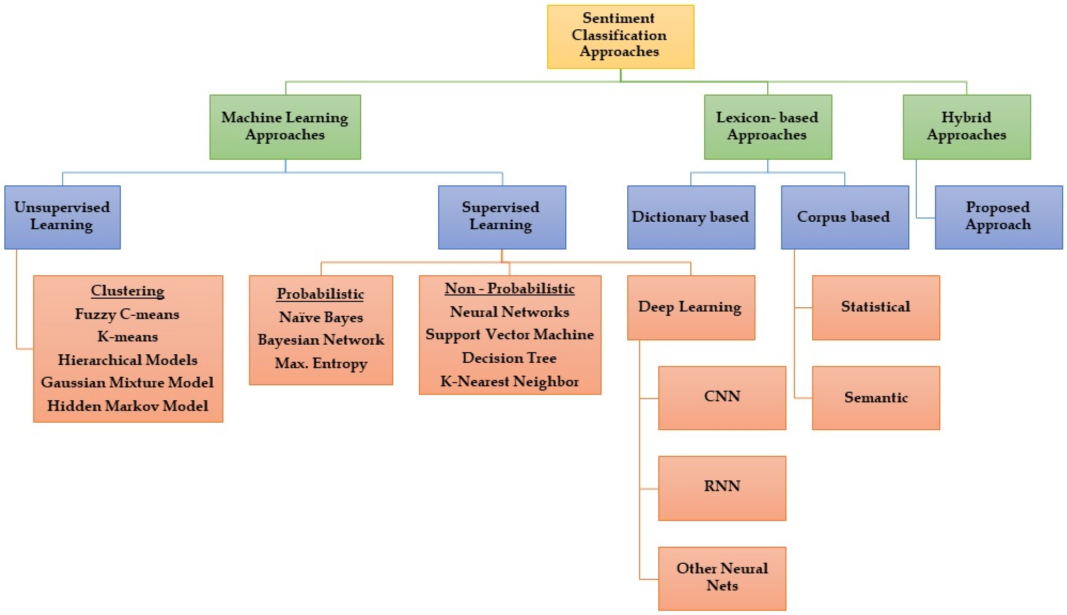

Figure 2: Text sentiment classification approaches.

and FastText models, to classify sentiment on a Twitter dataset using support vector machine (SVM). Ahuja et al. [52] investigated the effects of TF–IDF and n-gram on SA. The author Mohey El-Din [53] proposed an upgraded BOW to analyse textual reviews from the CiteULike website. Mee et al. [54] used regression and SA, specifically TF–IDF, to investigate the relationship between textual properties and Twitter user characteristics.

3.1.3 TEXT SENTIMENT CLASSIFICATION APPROACHES

SA is a promising research field with application in various fields. As a result, researchers are always proposing, evaluating and comparing new methods to improve SA performance. Existing SA methods may be classified into four categories: ML, lexicon-based, hybrid and DL techniques. The most extensively used method is ML, which uses ML methods and linguistic properties to classify sentiments. The Lexicon-based approach utilises a pre-defined sentiment lexicon to define the sentiment score for a document by integrating the sentiment scores of all the terms in the document. The pre-defined lexicon contains a set of frequently used words or phrases with their corresponding sentiment score. Hybrid methods combine machine learning and lexicon-based approaches to address the weaknesses of each methodology and improve sentiment classification performance. DL algorithms use word embedding to extract the most relevant properties from the text, outperforming ML-based techniques for textual sentiment classification. This section briefly summarises the textual SA approaches shown in Figure 2. The most notable advantages and disadvantages of these approaches are listed in Table 1.

3.1.3.1) MACHINE LEARNING APPROACHES

Machine-learning approaches are one of the most widely utilised techniques for categorizing textual information into different sentiment labels (e.g. negative, positive and neutral) using training and testing parts of the textual dataset. These techniques begin with the creation of a training set and labelling it with the sentiment. After a statistical analysis is performed on the training data, a collection of features is retrieved and sent to a classifier model. Once the classifier has been trained with sentiment labels, it will be able to estimate the sentiment polarity of an unannotated sample [55]. According to [56], these methods can be classified as supervised [57] and unsupervised [58] learning techniques. The supervised strategy is used when the classification problem involves a specified set of classes. However, the unsupervised approach may prove the optimal solution when it is impossible to establish this set due to a lack of labelled data. Although ML approaches may recognize domain-specific structures in text, which improves classification performance, they usually require massive training datasets to achieve good performance. Moreover, the classifiers developed for one dataset may underperform in another [59, 60].

Table 1: Advantages and disadvantages of various sentiment analysis approaches

Method	**Advantages**	**Disadvantages**
ML	• Dictionary is unnecessary • Excellent classification accuracy • High precision and adaptability	• More time is required • Controlled by the domain • Requires human interaction and information labelling • Does not consider the sentiment information encoded in sentiment terms
Lexicon-based	• No labelling is required • Depends on the domain • Takes less time	• Requires strong language resources • Low precision • Requires dictionaries with various viewpoints • Ignore the sentence context
Hybrid Method	• Takes less time • Combines the advantages of several procedures • Can identify and quantify sentiment at the concept level • More accurate	• Low precision • Lacks reliability
DL	• Features are automatically identified and optimised • Same NN may handle several tasks and data types • DL architecture is more adaptable to future issues • It captures the nonlinear interaction between the user and the object, allowing for faster sentiment prediction	• Requires a large amount of labelled training samples to achieve good performance • Expensive training due to complex data models • Long training time

3.1.3.1.1) SUPERVISED LEARNING

Supervised learning is established by training the system using a labelled dataset, with the labels denoting the classes (e.g. positive, neutral and negative). The classes are given several comparable sets of features and distinct labels. To categorise the training data, supervised learning approaches use probabilistic or non-probabilistic classification models. The probabilistic classifier makes a prediction based on the probability distribution across classes and calculates the likelihood that the extracted features belong to one of the classes (0 or 1). It often uses Bayes' theorem as its foundation [61]. The classifier uses mixture models to conduct classification, with each class being an element of the mixture. Probabilistic classifiers are easy to build, computationally efficient compared with other techniques and require little training data. However, if the data do not completely match the distribution assumptions, then classification performance may suffer. In certain cases, the probabilistic classifier may be inadequate. The disadvantages of using a probabilistic classifier may be avoided by utilising a non-probabilistic technique. A non-probabilistic classification approach does not exhibit class propagation and is useful when probabilistic classifiers are ineffective. It distinguishes the feature space from the sample origin and returns the class associated with that space. The subsections that follow give an overview and comparison of the most commonly used supervised classification algorithms for SA.

Naïve Bayes (NB). It is one of the most frequently used probabilistic text classifiers. The model is based on Bayes' theorem for forecasting the probability of a given set of characteristics, and the input features are assumed to be naturally self-regulating. The overall probability is determined by multiplying the prior probability by the likelihood. This method simplifies the training and classification processes. Z. Li et al. [62] developed a danmaku sentiment dictionary and proposed a novel approach for analysing the sentiment of danmaku reviews using a sentiment dictionary and NB. The approach is useful for monitoring the broad emotional orientation of danmaku videos and forecasting its popularity. Shrivastava et al. [63] developed a Twitter SA programme to analyse Republic of Indonesia's presidential candidates in 2019. This SA method involves data collection using Python libraries, text processing, data training and testing. Finally, text classification is performed using an NB approach. Kunal et al. [64] suggested utilising Tweepy and TextBlob, a Python framework, to access and classify tweets through NB. Kharisma et al. [65] used term weighting to obtain the best model combination. The tagged data were cleaned and transformed into structured data that can be analysed. The pre-processed data were then categorised using multinomial and Bernoulli NB. The experiments showed that Bernoulli NB is more accurate than multinomial NB.

Maximum Entropy (ME). It is a probabilistic classifier that is mostly used for text classification based on distribution uniformity. It depends upon the assumption of known data and ignores unknown data. This technique uses search-based optimisation to estimate the weights of the characteristics to predict the lowest risk while maximising entropy. Habernal et al. [66] conducted a performance evaluation of multiple classifiers to assess sentiment polarity, revealing that the ME classifier outperforms all other classifiers in predicting sentiment polarity on a given dataset. Htet and Myint [67] created a social media (Twitter) data analysis system to determine the state of people's health, education and business based on Twitter data, in which SA is utilised to propose these specified requirements using the ME classifier. On the basis of ME and NB, Ficamos et al. [68] developed a sentiment classification approach that uses PoS tags to extract unigram and bigram features. They obtained greater estimation accuracy by concentrating on a certain issue. Hagen et al. [69] advocated combining SVM with other models, such as ME and stochastic gradient descent optimisation, in comparison with previous SemEval editions with different feature sets, the combined model were selected as the best performing model of the year.

Bayesian Network (BN). It is a probabilistic graphical model that depicts the relationship between random variables. The distribution of this model may be described simply as a directed acyclic graph, with vertices and edges representing variables and dependence relations, respectively. The network's structure is easily expandable, which allows for the addition of new variables. BN can help with the decision-making process of a complicated problem by estimating the probability of its causes and consequences. Gutierrez et al. [70] studied the literature on the use and significance of BN in the SA area to assess textual emotions as part of research on text representation and BN. Wan et al. [71] suggested an ensemble sentiment classification technique based on the majority vote principle of several classification methods, including NB, SVM, BN, C4.5 decision tree and random forest algorithms, to classify tweets about airline services.

Support Vector Machine (SVM). It is a linear classifier that can manage discrete and continuous variables and divide data linearly or nonlinearly. It has a strong theoretical foundation and outperforms most other algorithms in terms of classification accuracy. The primary objective of an SVM classifier is to identify the optimal hyperplane for class separation. Hyperplanes with a large margin to a training point from either class are more efficient because they minimize the generalisation error of a classifier.

SVM has been used successfully in a number of investigations. Al-Smadi et al. [72] presented state-of-the-art methodologies that rely on supervised ML for dealing with the difficulties of Arabic hotel ratings. Deep recurrent NN (RNN) and SVM are developed and trained by utilising lexical, word, syntactic, morphological and semantic information. A review dataset of Arabic hotels has been used to evaluate the proposed methodologies; the SVM strategy outperforms the other deep RNN approach in the evaluated task. Nafis et al. [73] suggested an improved hybrid feature selection strategy to improve ML sentiment classification. Firstly, two customer review datasets were obtained and pre-processed. Next, they introduced a hybrid feature selection framework that uses TF–IDF and SVM–recursive feature elimination to assess and rank the features iteratively. Both approaches were developed and tested using the two datasets, and only the k-top features from the ranking features were utilised for sentiment classification. Finally, the proposed approach was evaluated using SVM classifier. Obiedat et al. [74] proposed a hybrid strategy integrating SVM, particle swarm optimisation and other oversampling algorithms to deal with unbalanced data. SVM has been utilised as a ML classification approach by improving the dataset of reviews from several restaurants in Jordan. Rezwanul Huq et al. [75] proposed two strategies to properly categorise the sentiment label from tweets, one based on k-nearest neighbour and the other on SVM. Naz et al. [76] developed a technique that deals with Twitter sentiment classification using n-gram and SVM classifiers. They investigated the influence of weighting on classifier accuracy using three alternative weighting methods. They also used a sentiment score vector of tweets to enhance the performance of the SVM classifier.

K-Nearest Neighbour (KNN). It is a non-probabilistic classifier that is fairly easy to implement and works by comparing the new data points with the training points. It computes the Euclidean distance between the new point to its neighbours and assigns it to one of the classes based on the distance. Here, K represents the number of neighbours, and the new data point is then assigned to the class that contains the majority of its neighbours. However, if the number of independent variables exceeds a certain threshold, the algorithms become difficult. Naw Naw [77] conducted sentiment classification on education, business, crime and health using KNN and SVM classifiers. The system provided educational authorities, economists, government organisations and health analysts with results related to these domains. Rizzo Irfan et al. [78] conducted an emotional analysis of the 2013 curriculum. The KNN approach was used for sentiment classification. An ensemble of numerous feature sets was used, including textual, Twitter-specific, lexicon-based, PoS and BOW features. The experiment results revealed that ensemble features outperform individual features when it comes to sentiment classification. Damarta et al. [79] employed text mining to regulate the service quality in PT PLN (Persero) by categorising Twitter data using KNN. The obtained data were pre-processed and categorised as negative, neutral or positive. The results achieved 87.41% accuracy using KNN. Muktafin et al. [80] utilised KNN and TF–IDF algorithms with NLP to categorise conversations into ‘satisfied’ and ‘dissatisfied’ categories. Their results achieved 74% accuracy.

Decision Tree (DT). This technique uses a set of if–then rules to construct a tree-like structure with nodes and branches that are linked together, making it suitable for regression and classification problems. This network has a root node, decision nodes and leaf nodes. Decisions are made at decision nodes depending on the characteristics extracted from the dataset. DTs work admirably even when the input data volume is large. Phu et al. [81] categorised positive, negative and neutral semantics for English texts using an ID3 approach of a DT. The semantic classification was based on several criteria produced by the ID3 algorithm on 115,000 English texts. The test was created using 25,000 English texts, and 63.6% accuracy was obtained. Es-Sabery et al. [82] developed a unique MapReduce-enhanced weighted ID3 DT classification algorithm for Opinion Mining, which includes three aspects. Firstly, n-gram or character-level, BOW, word embedding (GloVe and word2vec), FastText and TF–IDF feature extractors were used to effectively discover and collect important features from the provided tweets. Secondly, feature selections (e.g. chi-square, gain ratio, information gain and Gini index) were investigated to reduce the dimensionality of the high features. Thirdly, the characteristics were classified using an enhanced ID3 DT classifier that uses weighted information gain rather than the standard ID3 information gain. Fitri et al. [83] conducted SA in three stages: data pre-processing, classification, and evaluation, in order to determine the sentiment polarities (positive, negative, or neutral) in people’s reviews on an anti-LGBT campaign in Indonesia. They developed different ML algorithms, including NB, DTs and RF, to perform the classification. The experiments showed good results related to each classifier.

Neural Network (NN). It has lately gained popularity as a useful categorisation approach. It works by extracting features from linear datasets and modelling the output as a nonlinear function of these features. A common NN architecture consists of three layers: input, output and hidden layers, each containing many ordered neurons. The connection between two consecutive layers is provided through the links between the neurons of each layer. In a gradient descent training procedure, each link has a corresponding weight value that is calculated by minimising a global error function. There may be several hidden layers with a single input and output layer, where the number of output classes equals the number of output nodes. The feedforward NN is one of the most frequently used NNs, in which signals travel in one direction, from the input to the hidden layers and finally to the output layer. Bhargava et al. [84] suggested a method for analysing tweets in one of the Indian languages (Hindi, Bengali or Tamil). To avoid overfitting and error accumulation, 39 sequential models were constructed with optimal parameter values using RNN, long short-term memory (LSTM) and CNN. Moraes et al. [85] proposed document-level sentiment classification to automate the classification of textual reviews on a particular subject as positive or negative. A standard BOW technique was used to extract all the important features, and the classification was performed using SVM, NB, and artificial neural network (ANN). In the experiments, the ANN outperformed the SVM on the benchmark dataset of movie reviews. Vinodhini and Chandrasekaran et al. [86] compared NN-based sentiment classification approaches (i.e. backpropagation NN, probabilistic NN (PNN) and homogeneous ensemble of PNN) employing varying degrees of word refinement as features for feature-level sentiment classification. The outcomes of ANN-based approaches were compared with two statistical methods using a dataset of Amazon product reviews. In classifying the sentiment of product reviews, the PNN outperformed the other two NN algorithms. The use of NN-based sentiment classification combined with principal component analysis as a feature reduction methodology shortens the training time. Table 2 details the advantages and disadvantages of probabilistic (i.e. NB, BN and ME) and non-probabilistic (i.e. SVM, KNN, DT and NN) classifiers.

3.1.3.1.2) UNSUPERVISED LEARNING

The fundamental limitation of supervised SA is the need for large annotated datasets to train a classification algorithm. Domain adaptation of supervised classifiers is difficult because certain domains need a more formal or lengthier text input (e.g. film reviews) that help in building the treebanks dataset. The same approach cannot be used for social media data due to brief reviews (tweets are 140 characters in length). When the labelled data is unavailable, unsupervised SA is used. This approach does not require labelled data to produce sentiment predictions, which reduces the labelling costs. Unsupervised clustering finds a structure in the incoming data point and groups it with similar objects/points. Data points are assigned membership values via the fuzzy C-means method based on their distance from the cluster centre. A higher membership grade signifies closer proximity to the cluster's core. In K-means clustering, the clusters have K randomly chosen centres. Each input data point is allocated to the nearest centre, and new centres are calculated. The algorithms finish when K does not change.

Many studies have used unsupervised approaches. Harish et al. [87] proposed a new method based on term frequency vectors using clustering to encode text corpora. The mean and standard deviation of each cluster's phrase frequency vectors were expressed symbolically (interval-valued). The term frequency vectors were clustered using a fuzzy C-means clustering algorithm based on an adaptive squared Euclidean distance between interval vectors. They used popular datasets (e.g. 20 Newsgroup Large, 20 Mini Newsgroup and Vehicles Wikipedia) and innovative datasets (e.g. Google Newsgroup and Research Article Abstracts) to evaluate the suggested model's performance. The results revealed that acquired classification accuracy outperforms existing methods. Orkphol and Yang [88] used K-means to group similar microblog tweets indicating relevant sentiments about a product. Given the nature of microblogging, the dataset was sparse and high-dimensional. To solve this challenge, the authors used TF–IDF to choose important characteristics and singular value decomposition to decrease the high-dimensional dataset while keeping the most relevant features. The artificial bee colony (ABC) was used to determine the ideal starting state of centroids for K-means, the silhouette analysis method was used to determine the perfect K and SentiWordNet was used to score each group after categorising them into K groups. Their technique showed its effectiveness in improving the K-means results. Han et al. [89] suggested a semi-supervised based on a dynamic threshold and multiple classifiers to manage the shortage of initial annotated data. The training data were repeatedly auto-labelled using a dynamic threshold function. Moreover, the proposed weighted voting approach compared the performance of various SVM classifiers. The proposed model achieved maximum SA accuracy across datasets of various initial labelled training data sizes. Fernández-Gavilanes et al. [90] constructed an unsupervised dependency parsing-based text classification system that uses sentiment lexicons and NLP approaches. The Cornell Movie Review, Obama–McCain Debate and SemEval-2015 datasets showed the 'resilience and efficiency of the system. García-Pablos et al. [91] described a W2VLDA system that performs aspect category classification, aspect-term and opinion-word separation and sentiment polarity classification for any domain or language. They used the SemEval 2016 task 5 (ABSA) dataset to assess the domain aspect and sentiment classification. Many industries (e.g. hotels, restaurants and electronics) and languages (e.g. English, Spanish, French and Dutch) had competing outcomes.

3.1.3.1.3) LEXICON-BASED APPROACHES

In SA, lexicon-based approaches are one of the most basic methods to the study of sentiment. It uses an opinion lexicon (a predetermined set of words) that utilises scores to categorise words as negative or positive. A score may be a basic polarity value (such as +1, -1 or 0) for positive, negative and neutral terms, or a numerical value expressing emotional power or intensity. The semantic orientation values assigned to a document's words determine its final orientation. When documents are tokenized into small words or short phrases, emotional scores are assigned to each element from the lexicon. The final sentiment can be calculated by summing or averaging each word's emotional scores. The lexicon-based technique works well for phrases and features. It is unsupervised because it does not require training data. The fundamental issue of this strategy is domain dependency because words may have several meanings and senses, making a good term terrible in another. This issue can be solved by creating a domain-specific sentiment lexicon or modifying an existing vocabulary. 'The three major methods of collecting opinion words are manual-, dictionary- and corpus-based approaches. Manual procedures are time-consuming and are not always used. It is typically used as

Table 2: Advantages and disadvantages of supervised learning classifiers.

Method	Advantages	Disadvantages
NB	• Simple to implement and understand • Low-resource computing demands • Less data and training time than other approaches	• Assumes characteristics are independent, which is seldom the case • Restricted by data scarcity because a probability value must be determined for each possible value.
BN	• Less time to build a model because it is understandable even in complicated domains • Manages missing data and gets accurate results with limited training data • Prevents overfitting	• Computationally costly; thus, it is rarely used • Unsuitable for problems that contain many features
ME	• Useful when the previous distributions are unknown • Quick at obtaining information from text and dealing with massive amounts of data	• Has a tendency towards overfitting • Lacks reliability
SVM	• Stable in high-dimensional spaces • More accurate and simpler to train • Memory-efficient due to high-dimensional kernel mapping • Strong and can manage large, sparse collections of samples	• Insufficient performance when the number of features exceeds the samples number • Requires a proper kernel function • Poor interpretability due to lack of probabilistic explanation
KNN	• Simple to comprehend and apply • Trains quickly • Handles noisy data well • Works well with samples that contain several class labels	• When comparing unlabelled samples to a large number of possible neighbours, lazy learners suffer high computing costs • Sensitive to data local structure • Limited memory • Operates slowly due to its supervision
DT	• Easy and quick • Yields accurate results • Understandable representation • Facilitates gradual learning • Saves memory • Handles noisy data • Finds the optimal split attribute using entropy, Gini index and information gain	• Requires extensive training • Due to replication issue, DTs may have considerably more complicated representations for some ideas • Overfitting issues
ANN	• Handles complicated relationships between variables and greater generalisation even with noisy data • Efficient for situations with high dimensionality • Fast execution	• Theoretically complicated and difficult to execute • High memory consumption • Requires more training time and a huge dataset in certain situations

a last check after the other two automated processes to verify no errors are made. According to the dictionary-based approach, synonyms have comparable emotional polarities, whereas antonyms have opposing emotional polarities. This method generates emotion lexicons utilising WordNet [92] and the thesaurus [93]. It starts by manually gathering known-orientation seed words. Synonyms and antonyms are then searched for each word in the list. The list is iterated until no new words are found. Some errors necessitate human intervention. Different from dictionary-based approaches, corpus-based methods start with a list of seed sentiment words that have a predefined orientation and then use syntactic or co-occurrence patterns to find more emotional words that match their orientation in a large corpus. It includes a statistical method for determining a word's polarity based on how often it occurs in negative text and a semantic method for determining sentiment based on word similarity. Table 3 lists the advantages and disadvantages of the dictionary- and corpus-based approaches.

Many researchers have used lexicon-based approaches. M. Huang et al. [94] developed sentiment CNN to analyse sentences' emotions using contextual and sentiment information from sentiment terms. Word embeddings give contextual data, whereas lexicons provide sentiment information. A highway network is created to allow for the adaptive integration of sentiment and contextual information from sentences by increasing the association between their characteristics and sentiment words. Yu et al. [95] used an alternative sentiment embedding learning method to enhance the current pre-trained word vectors by building a model for word embedding refinement that uses real-valued sentiment intensity ratings from sentiment lexicons. Jurek et al. [96] introduced a new lexicon-based SA method for real-time tweet content analysis. The approach comprises two core parts: sentiment normalisation and an evidence-based mixture function, which are used to forecast sentiment intensity instead of positive/negative labelling and promote the usage of the combination of the sentiment classification procedure. Sanagar and Gupta [97] developed a robust unsupervised sentiment lexicon learning technique for new genres. The method and progressive learning are used to extract polarity seed concepts from a corpora of randomly selected source areas. Thereafter, the genre-level information is delivered to the target areas. Unlabelled training data from the source and target areas are utilised to construct a sentiment lexicon using latent semantic analysis. In a dictionary-based approach, Park and Kim [98] developed a dictionary-based strategy to create a thesaurus lexicon designed for sentiment classification. The proposed approach gathered a thesaurus utilising three online dictionaries and only stored co-occurrence keywords to enrich the reliability of the thesaurus lexicon. Also, this approach expanded the thesaurus vocabulary, which is a collection of synonyms and antonyms, to improve post accessibility and sentiment classification performance. Sanagar and Gupta [99] presented a survey paper discussing the polarity lexicon based on two parts; first is a review of the literature from ancient to current methodologies, and second is a discussion of free-source polarity lexicons. Wang et al. [100] developed an improved random subspace method called PoS-RS, a sentiment classification method based on parts of speech that regulates the balance between accuracy and variety of base classifiers by including two essential factors, namely, content and function lexicon subspace rates. Jha et al. [101] constructed an emotion-aware dictionary using various domain data, including annotated data from the source domain and unlabelled data from the source and target domains. It is then used to identify unlabelled reviews in the target domain. The recommended strategy identifies 23%–24% more terms in the target domain than existing approaches. In corpus-based approach, Agarwal and Mittal [102] proposed a method for SA using corpus-based semantic orientation. They investigated unique sentiment-bearing feature extraction strategies and conventional techniques, such as unigram, PoS-based and dependency features. In addition, they provided an approach for creating SA methods for areas with sparsely labelled data. Luo et al. [103] developed a novel recommendation technique that incorporates the sentiment aspect of tagging into a social tag recommendation system. The tag sentiment data, which are provided in the form of users' subjective polarity towards annotated resources, function as an additional information filter to offer users more desired and positive resources. Thus, sentiment tagging can improve recommendation efficiency.

3.1.3.1.4) HYBRID APPROACHES

The hybrid technique combines lexicon- and ML-based approaches. It is principally inspired by the need to combine ML's high precision and flexibility, with the robustness of a lexicon-based approach to deal with ambiguities and incorporate the context of sentiment words. Thus, the hybrid strategy may improve sentiment classification by combining both methodologies to overcome their disadvantages and maximise their advantages [104]. The true value of this technique is the presence of the lexicon/learning symbiosis, which can be utilised to enhance the efficacy and usability of the lexicon. It achieves high model accuracy by using strong supervised learning methods. The drawback is that the ideas expressed include many irrelevant terms related to the subject. Due to this noise, feelings are often assigned a neutral polarity rather than being detected as positive or negative [105]. Ghiassi et al. [106] created a unique hybrid approach by combining dynamic ANN with n-grams. They constructed two classifiers, an SVM and an ANN, using emoticons and tweets as features. Khan et al. [107] suggested a hybrid method to twitter feed categorization. The suggested technique involves pre-processing the text before providing it to the classifier. Devi et al. [108] classified documents using ML models that combined dictionaries and the HARN algorithm presented in a lexicon-based approach. The reviews were domain-classified using NB and SVM classification algorithms, and polarity was then calculated at the document level using the HARN technique. The result showed that the hybrid method is better than the HARN

Table 3: Advantages and disadvantages of dictionary- and corpus-based approaches

Method	Advantages	Disadvantages
Dictionary-based Approach	• Computationally affordable if dataset training is not required • Represents a solid technique to quickly develop a vocabulary with many sentiment words and their orientation	• Fails to discover sentiment words with domain particular orientation and is thus unsuitable for context- and domain-specific categorisation • Compiling dependency rules is time-consuming
Corpus-based Approach	• Simplicity	• Requires a large dataset to identify word polarity and hence textual sentiment • Depends heavily on the polarity of the words in the training corpus, because polarity is calculated for the terms in the corpus

algorithm by 80%–85%. Shin et al. [109] developed a CNN that incorporates lexicon embeddings and an attention mechanism. They made lexicon embeddings by merging word scores from several lexicons. Three methods were utilised to create a CNN model: nave concatenation, multichannel and separate convolution. They showed that lexicon integration improves the performance, reliability and effectiveness of CNN. Elshakankery et al. [110] suggested a hybrid incremental learning method for Arabic tweet SA (HILATSA), which uses ML and a lexicon-based approach, for determining sentiment polarity in tweets. They classified and created a lexicon of words, emoticons, idioms and other important lexica using SVM, logistic regression and RNN. They studied the Levenshtein distance in SA to handle different word forms and spelling issues. HILATSA was tested using six datasets, the results found that combining ML and lexicon-based methodologies in textual SA for social networks is effective.

3.1.3.1.5) DEEP LEARNING APPROACHES

The implementation of ANN-based DL has considerably enhanced the SA field. DL is a rapidly growing topic in ML that provides techniques for comprehending feature representation in a supervised or unsupervised manner [111]. DL is a concept that refers to NNs with numerous layers of perceptron motivated by the human brain [112]. This architecture allows for the training of more complicated models on a considerably larger dataset, resulting in an outstanding performance in a wide variety of application domains, including computer vision, speech recognition and NLP [113].

DL comprises various NN models, including CNN, RNN and others that have been adopted in many research studies. Alayba and Palade [114] improved sentiment classification by combining CNNs and LSTM networks and removing CNN's max-pooling layer, which reduces the length of the generated feature vectors by convolving the filters on the input data. As a result, the LSTM networks will receive vectors that have been sufficiently captured by the feature maps. The authors also evaluated many successful strategies for developing and expressing text features to increase the accuracy of Arabic sentiment classification. Salur and Aydin [115] proposed a unique hybrid DL model that strategically integrates several word embeddings (i.e. word2vec, FastText and character-level embeddings) with various DL algorithms (i.e. LSTM, gated recurrent unit (GRU), bidirectional LSTM (BiLSTM) and CNN). The proposed model combines features from several DL word embedding techniques and classifies texts according to their emotion; it achieves better result in terms of sentiment classification. Zulqarnain et al. [116] developed a unique two-state GRU and encoder approach called E-TGRU to construct an effective framework for SA. The findings showed that given sufficient training data, The GRU model is capable of effectively acquiring the words used in user opinions. The Results presented that E-TGRU outperformed GRU, LSTM and bi-LSTM using Internet Movie Database (IMDB) and Amazon Product Review datasets. L. Li et al. [117] proposed an SA model for online restaurant reviews by integrating word2vec, bidirectional GRU and the attention technique. The findings showed that the model outperformed other popular SA models in terms of overall performance. Sharma et al. [118] suggested word2vec word embedding and CNN for classifying short sentences. The method uses a pre-trained word2vec model to create word vectors and a CNN layer to extract improved features for sentence classification. Basiri et al. [119] proposed two deep fusion models based on a three-way decision theory to investigate drug reviews. When the deep method achieved low confidence in classifying the test samples, the first fusion model (3W1DT) was developed using a DL method as the main classifier and a traditional learning method as the secondary classifier. In the second fusion model (3W3DT), three extremely deep models and one traditional model were trained on the whole training set, and each model classified the test sample independently. The drug review test sample was then classified using the most confident classifier. The findings indicated that the 3W1DT and 3W3DT approaches outperformed standard and DL methods. W. Li et al. [120] introduced a novel padding approach that resulted in a more stable size of the input data sample and an increase in the quantity of sentiment information in each review. By using parallelisation, a DL-based SA model called lexicon was integrated with two-channel CNN–LSTM/BiLSTM family models. The experiments on various difficult datasets, such as the Stanford Sentiment Treebank, showed that the proposed strategy outperforms a broad range of baseline techniques. Abid et al. [121] designed a unified architecture for sentiment classification on Twitter that integrates an RNN model for extracting long-term dependencies with CNNs and GloVe as a word embedding approach. The experiments outperformed the baseline model on the Twitter corpora. Dang et al. [122] proposed hybrid deep SA learning models that integrate LSTM networks, CNNs and SVMs and were designed and evaluated on eight textual tweets and review datasets from various areas. On all datasets, the hybrid models outperformed the single models in SA. Fatemi and Safayani [123] proposed a generative framework for joint sentiment topic modelling based on NNs by changing the restricted Boltzmann machine structure and adding a layer corresponding to the sentiment of the text data. A contrastive divergence algorithm was used to manage and implement the proposed method. The new connected layer in the proposed method was a multinomial likelihood distribution layer, which could be used in textual sentiment classification or any other application.

3.1.4 TEXTUAL DATASETS

Researchers on SA might use their data or datasets that are publicly available. In order to be relevant to a certain problem, many researchers acquire new datasets relating to the subject. The main drawback is that the dataset must be labelled, which takes time. Furthermore, generating a large amount of data is not always simple. The availability and accessibility are the primary factors for selecting the datasets. Table 4 highlights the most popular datasets used in textual SA.

3.2 VISUAL SENTIMENT ANALYSIS

Popular social platforms, such as Flickr and Instagram, provide considerable amounts of visual data in the shape of images. VSA recognises and extracts sentiment and emotion from facial expressions or body movements. In addition to objects or entities, activities and locations, visual materials may include specific indicators that transmit attitude and emotion signals. For example, an image of a tasty meal, a gorgeous landscape or a lovely ceremony is likely to convey the publisher's optimistic attitude. After condensing these sentimental experiences into semantic labels, we can construct computer vision challenges to uncover useful mappings from low-level visual information (e.g., raw pixels and motions) to high-level emotion labels in classification, localisation and summarisation tasks. Throughout history, several studies have used image analysis to focus on specific cognitive and psychological applications, such as human face recognition detection and social marketing campaigns. This section describes the various image pre-processing operations, feature extraction methods and computational methodologies used in VSA.

3.2.1 VISUAL PRE-PROCESSING

Image pre-processing represents the fundamental processing of pictures. It aims to modify the image data by removing unwanted distortions or enhancing particular visual properties that are important for further processing and analysis. Some of the most often used pre-processing operations, namely, pixel brightness transformation (PBT)/brightness correction, geometric transformations, image filtering and segmentation and Fourier transform and image restoration, are discussed as follows.

3.2.1.1) PBT

Brightness transformations improve the pixel brightness, and the transformation is determined by the properties of the pixel. In PBT, the value of an output pixel is exclusively determined by the value of the corresponding input pixel. Such operators include brightness and contrast changes, as well as colour correction and transformations. Contrast enhancement is an important component of image processing for human and computer vision. It is often utilised in medical image processing, voice recognition, texture synthesis and various other image or video processing applications. Brightness transformations may be categorised into brightness corrections and grayscale transformations. The most frequent procedures for PBT are gamma correction or power law transform, histogram equalisation and sigmoid stretching.

- Gamma Correction. It is a nonlinear adjustment of the pixel value. Linear operations, such as scalar multiplication and addition/subtraction, are used on individual pixels in picture normalisation. Conversely, gamma correction uses a nonlinear operation on the source image pixels, which might change the image saturation.

Table 4: Popular datasets for textual SA

Datasets	Links
IMDB	https://datasets.imdbws.com/ https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
Stanford Sentiment Treebank	https://nlp.stanford.edu/sentiment/treebank.html
Cornell Movie Reviews	https://www.cs.cornell.edu/people/pabo/movie-review-data/
Thomson Reuters text research collection	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IEJ2UX https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection
NYSK dataset	https://archive.ics.uci.edu/ml/datasets/NYSK#
ABC Australia News Corpus	https://live.european-language-grid.eu/catalogue/corpus/5162
Sentiment labelled sentences dataset	https://www.kaggle.com/datasets/marklvl/sentiment-labelled-sentences-data-set
Sentiment 140	https://www.kaggle.com/datasets/kazanova/sentiment140
Tweets Airline	https://data.world/crowdflower/airline-twitter-sentiment
Tweets SemEval	http://alt.qcri.org/semeval2016/task4/ http://alt.qcri.org/semeval2015/task10/ http://alt.qcri.org/semeval2017/task4/
Book Reviews and Music Reviews	https://data.world/dataquest/book-reviews https://www.kaggle.com/datasets/cakiki/muse-the-musical-sentiment-dataset https://webscope.sandbox.yahoo.com/catalog.php?datatype=r

- Histogram Equalisation. It is a popular contrast enhancement method that works on almost any picture format. It is a complex approach for adjusting an image's dynamic range and contrast by redesigning its intensity histogram. Histogram modelling operators use nonlinear and non-monotonic transfer functions to map between pictures instead of linear and monotonic transfer functions.
- Sigmoid Stretching. A sigmoid function is a nonlinear activation function. The term 'sigmoid' comes from the S form of the function, which is called logistic function by statisticians.

3.2.1.2) GEOMETRIC TRANSFORMATIONS

A geometric transformation modifies the positions of pixels in an image while keeping the colours the same. It enables the reduction of geometric distortion that occurs during the acquisition of an image. The standard geometric transformation operations are rotation, shearing, translation, scaling, distortion (or un-distortion), Affine Transformation which combines the four transformations and Perspective Transformation which alters the perspective of an image or video to gain a better understanding of the required information.

3.2.1.3) IMAGE FILTERING AND SEGMENTATION

Filters are used to improve the visual qualities of images and/or extract important information, such as edges, corners and blobs. The kernel of a filter is a small array applied to each pixel and its neighbours in an image. Some of the basic filtering techniques are as follows.

- Low-pass Filtering (Smoothing). It serves as the foundation for most smoothing methods. Smoothing a picture involves averaging nearby pixels to reduce the variance between pixel values.
- High-pass Filtering (Edge Detection, Sharpening). It makes an image look sharper. Different from low-pass filters, high-pass filters enhance small details in the picture. It is similar to low-pass filtering except that it uses a different type of convolution kernel.
- Directional Filtering. The first derivatives of a picture may be calculated using directional filtering, which is an edge detector. When the values of surrounding pixels shift considerably, the initial derivatives (or slopes) become particularly visible. Any spatial direction may be filtered with a directional filter.
- Laplacian Filtering. A Laplacian filter is an edge detector that measures the second derivatives of an image by measuring the rate at which the first derivatives change. It evaluates whether a change in the values of neighbouring pixels is caused by an edge or is part of a continuous development. Negative values are often included in Laplacian filter kernels as a cross pattern centred within the array. The corners may be zero or positive, and the centre may be positive or negative.

3.2.1.4) IMAGE SEGMENTATION

Image segmentation is a common technique in digital image processing and analysis for dividing an image into segments or areas based on pixel attributes. It may separate the foreground and background pixels or cluster pixels based on colour or shape. The three types of image segmentation method are as follows.

- Non-contextual Thresholding. Thresholding is the easiest non-contextual segmentation approach. With a single threshold, this algorithm turns a greyscale or a coloured picture into a binary region map with a single threshold.
- Contextual Segmentation. Non-contextual thresholding combines pixels regardless of where they are on the image plane. Contextual segmentation can better identify various objects because it considers the closeness of pixels that belong to an object.
- Texture Segmentation. Many image analysis and computer vision applications rely on texture. It divides an image into areas with different textures that use comparable sets of pixels.

3.2.1.5) FOURIER TRANSFORM AND IMAGE RESTORATION

Fourier transform is used to divide an image into sine and cosine components. The transformed image is in the Fourier or frequency domain, whereas the original image is in the spatial domain. Each point in the Fourier domain picture represents a frequency in the spatial domain image, Fourier transform is utilised in image analysis, filtering, reconstruction and compression.

Many researchers have used different types of pre-processing operations in their studies. J. Zhang et al. [124] proposed a method for detecting salient objects that immediately generates a limited set of detection windows for an input image. They created the silent objects' proposals using a CNN model and proposed an optimised method for generating a limited set of detection windows from the noisy proposals. Navaz et al. [125] used single-image super-resolution and DL methods to increase the resolution of images in an emotion detection dataset. It is a significant pre-processing method that uses CNN to create high quality images. Priya et al. [126] employed image augmentation techniques, including scaling, rotation and translation, to pre-process images in a dataset. During scaling, the object borders were trimmed, and rotation detected the item in any orientation.

3.2.2 VISUAL FEATURE EXTRACTION

The most difficult aspect of constructing a VSA system, and of designing a data analysis technique in general, is selecting which data characteristics appropriately contain the information the system seeks to infer. Three important levels of semantics that are continually used for visual feature extraction are as follows.

3.2.2.1) LOW-LEVEL FEATURES

These features explain diverse visual phenomena in a picture, most of which are related to the colour values of the image pixels in some manner. They often involve general characteristics, such as colour histogram (CH), histogram of oriented gradient (HOG) and GIST. Previous research on VSA [127] revealed that low-level features, such as colours and textures, might be used to express an image's emotional influence. Low-level features are divided into global and local features. Table 5 highlights the most important low-level feature extraction methods used in the literature.

Table 5: Description of the most important low-level feature extraction methods in VSA

Low-level/Global Features		
Features	**Methods**	**Description**
Colour	CH	CH was introduced by Swain et al. [128]. It is an image retrieval method that visualises colour distribution within a picture. A CH picture can be represented by a one-dimensional feature vector. CH methods are location-insensitive due to translation, rotation and zoom invariance. They are suitable for images that are difficult to colour-divide automatically or require less pixel space.
	Colour Moments	It is a simple and effective colour characterisation method proposed by Stricker et al. [129]. Typical moments, which contain the mean, standard deviation, and skewness, can describe the average colour, colour variance and colour offset, respectively, and sufficiently define the colour distribution in pictures.
	Colour Coherence Vector	It is a development of the CH method that was defined by Pass and Zabih [130] for identifying the colour features of a picture. For each pair of pixels, the histogram is divided into two halves. If the continuous area occupied by certain pixels in the group exceeds a given threshold, then the pixels are coherent and this area is connected; otherwise, the pixels are incoherent, and this area is disconnected.
	Colour Set (CS)	CS, defined by Smith et al. [131], is considered a simplified version of the CH. It refers to a binary vector denoted by the symbol BM, an m-dimensional binary space, with each space axis corresponding to a single search m. When the colour m emerges, a CS, as a two-dimensional vector in the BM space, corresponds to the colour selection; c[m] = 1; otherwise, c[m] = 0. CSs may be formed instantly by specifying a specific threshold for the CH.
	Colour Correlogram	It was proposed by J. Huang [132] as an enhancement over CH to describe the colour distribution of an image, which not only can display the percentage of pixels of a certain colour in an entire picture but also indicate the spatial connection between distinct colour pairings.
Texture	Statistical Method: Grey-level Co-occurrence Matrix (GLCM)	It mainly uses the grey distribution to quantify textural features, including homogeneity, directionality and density of a picture. Using simple, robust, adaptable and easy-to-execute statistical techniques has many advantages. However, these approaches often require a large number of statistical computations. In terms of statistical analysis, GLCM defined by Mukherjee [133] and Savita et al. [134] is the most often used method.
	Model Method: Markov Random Field (MRF)	It uses model parameters to represent texture characteristics. It not only defines the randomness of the texture location but also the texture's overall regularity. MRF is a popular random field model. It is supported by a main theory [135] that specifies the grey value of any pixel in the picture in relation to the pixel values surrounding the image and the grey values of nonadjacent pixels.
	Structural Method: Texture Primitive (TP)	It is a texture analysis technique based on the TP hypothesis presented by T. Tan and A. Constantinides [136]. In TPs, a 2×2 window is the smallest unit for describing the picture texture. It is more appropriate to regular artificial texture; thus, the structural technique has substantial practical limitations.
	Signal Method: Wavelet Transform (WT)	WT is a representative signal approach introduced by Mallat [137] that is used for the feature extraction of image texture. It has a long history of being used in picture texture analysis. It decomposes the signal into a set of fundamental functions that can be created by changing the parent function.
Shape	Fourier Descriptor (FD)	FD is a commonly used shape feature descriptor that describes the shape feature of an image via Fourier transformation of the object boundary; it has properties of low computational cost, clarity and flexibility to transition from coarse to fine description [138].
	Hough Transform (HT)	It is a widely used technique proposed by Hough [139] for distinguishing geometric forms that share certain characteristics from a picture and for recognising lines that appear at different angles.
	Invariant Moment (IMs)	Hu [140] proposed IM based on algebraic invariants, which use the image's grey distribution moments to characterise the distribution of grey features.
	Zernike Moment (ZM)	ZM is an orthogonal complex moment in a polar coordinate space defined over the interior of the unit disc. ZM's form characteristic is not susceptible to noise, and its value is not redundant because the kernel of the ZM is a set of orthogonal radial polynomials. These moments may characterise the shape and detail of an image.
GIST descriptor	Global Image Descriptor	It is a type of global image features that is critical for recognising scenes in an image. It is calculated by convoluting the filter with an image at different scales. As a result, the high- and low-frequency repeated gradient directions of an image may be measured.
Low-level/Local Features		
Local Binary Pattern (LBP)		The LBP function is a binary representation of a texture analysis method that compares the value of the centre pixel to its neighbours [141]. LBP recognises pixels in a picture by thresholding their neighbourhood and translating the result to a binary value. The LBP texture operator is frequently used due to its discriminating power and operational simplicity.
HOG		HOG is a feature descriptor that calculates and counts the gradient direction histogram of the local area of the image for computing objects in computer vision and image processing.
Scale-invariant Feature Transform (SIFT)		SIFT identifies extreme points in a scale space and extracts position, scale and rotation invariance.
Speeded-up Robust Feature (SURF)		SURF is a fast version of SIFT. It detects feature points using the determinant value of the Hessian matrix and speeds up operations with integral graphs.
Bag of Visual Words (BOVW)		The BOVW model can be used for image classification or retrieval by treating image features as words. It is a vector of counted occurrences of a vocabulary of local image features.

Table 6: Description of the most important mid-level feature extraction methods in VSA

Mid-level Features	
Features	**Description**
Visual Sentiment Ontology (VSO)	VSO was proposed by Borth et al. [142]. One of its advantages is that it incorporates more classes in well-known visual ontologies, such as large-scale concept ontology for multimedia and ImageNet. It may be utilised in huge sentiment applications, such as microblog SA. VSO was first used to construct SentiBank, which is then used to predict the emotion of a picture. SentiBank is a library of trained concept detectors offering a mid-level visual representation. Thus, it is more conceptual than low-level characteristics; experimentally, it is excellent for describing image sentiment.
Sentribute Mid-level Attribute	Yuan et al. [143] developed a Sentribute method, which uses sentimental scene attributes as mid-level features. Four mid-level attributes were selected: 1) material, 2) action, 3) surface aspect and 4) spatial envelope. Finally, Sentribute provides 102 mid-level qualities that are easy to comprehend and ready to use. Sentribute is less strict than VSO because the four types of mid-level quality are specified individually, whereas VSO has become a prominent mid-level emotion feature.
Object-based Visual Sentiment Concept Analysis	Despite the fact that SentiBank is beneficial for large-scale VSO, it faces two difficulties. That is, 1) object-based notions must be localised, and 2) there exists ambiguity in the visual sentiment annotation. To address the aforementioned issues, T. Chen et al. [144] developed object-based visual sentiment concept analysis. It is a hierarchical system that decomposes difficult issues into object localisation and sentiment-related idea modelling.
Visual Sentiment Topic (VST) Model	This VSO-based algorithm has two major issues with topic image SA. Firstly, a VSO-based model cannot determine which ANP most closely fits the principal sentiment of an image. Secondly, microblog photos are important in the same subject, but a VSO-based model cannot fuse multi-image information in sentiment prediction. D. Cao et al.'s [145] VST model uses multi-image information in the same subject to generate VST, focusing on the important VSO in a picture.

3.2.2.2) MID-LEVEL FEATURES

These features include greater semantic information, being easier to comprehend and having stronger emotional links. Many of the previous works on VSA have used the 1200D mid-level representation provided by Borth et al. [142]. The advantage of using adjective noun pairs (ANPs) over nouns or adjectives is that a neutral noun may be transformed into a strong feeling ANP. Such paired notions are also more noticeable than adjectives alone. Table 6 highlights the most important low-level feature extraction methods used in the literature.

3.2.2.3) HIGH-LEVEL FEATURES

These features represent the semantic concepts depicted in the images at a high level. Such feature representation may be created using pre-trained classification algorithms or semantic embeddings [146]. Pre-trained CNN-based models (e.g. VGG16, VGG19 [147], Xception [148], Inception V3 [149] and Resnet [150]) are state-of-the-art image classification models that extract high-level semantic information from photos. These models are pre-trained using image net, a massive visual database dedicated to the research on visual object identification.

3.2.3 VISUAL CLASSIFICATION APPROACHES

Modelling, recognising and utilising sentiments expressed by facial or body gestures or sentiments associated with visual multimedia are the key research topics in VSA. As a result, researchers are continually proposing, evaluating and comparing new methods to improve SA performance and solve the challenges in the sector. An overview of the most frequently used classification techniques for VSA is presented in this section. The majority of extant methods fall into three categories: ML, DL and transfer learning techniques. ML is the most often used method; it largely relies on manually developed features to extract and categorise visual information using ML techniques. To increase visual sentiment classification performance, DL algorithms extract the most important visual characteristics without using manually created features and overcome the limits of ML approaches. The visual sentiment classification issue may be addressed via transfer learning, a common technique in computer vision, because it enables for the quick creation of correct models. Transfer learning is often demonstrated by using pre-trained models trained on a large ImageNet dataset to solve a problem similar to the one at hand. With the computational expense associated with training such models, models from the published literature (e.g. VGG, Inception and MobileNet) can be used. Numerous research has been conducted in this area. Afzal [9] developed an optimisation-based SVM model for automated VSA. The features of the input pictures are initially taken from the pre-trained ResNet-18's weighed FC8 layer, where the relief method is used to analyse the modified weight. The SVM classifier was optimised using a hybrid optimisation methodology called the Holoentropy Life Choice Optimisation algorithm, which combines the advantages of life choice-based optimization and cross-entropy methods. The model was evaluated using the Emotion-6 and Abstract Art photo datasets, and it showed a great performance. Instead of using SVMs, Jia et al. [10] proposed a specific graphical model for classification using the same colour characteristics. A total of 23k digital photos were used, including images of paintings with descriptors such as pretty, casual, romantic and jaunty. Machajdik and Hanbury [127] investigated the features that help with visual classification. The most essential visual features were selected based on emotional responses to colours and art. Emotional pictures were categorised by the writers' eight emotional output (i.e. awe, anger, amusement, contentment, excitement, disgust, sad and fear). Amencherla and Varshney [11] investigated the association between the visual content aspects of Instagram images and the psycholinguistic mood of their hashtag captions. Many colour attributes (e.g. hue, saturation and value), as well as colourfulness, hue and colour warmth, were used to predict visual emotion in thousands of photos. The findings support and clarify several psychological assumptions about the relationship between colour and mood/emotion, such as the relationship between colour and pleasure. L. Wu et al. [12] developed a VSA method that considers global and local data. Emotion was first deduced from the entire collection of photographs. Secondly, whether significant elements exist in a photograph was assessed. Sub-images were cropped from the whole image based on the detection window of the notable elements, if there are any. Furthermore, a CNN model was trained for the group of sub-images. The final results were obtained by combining sentiment predictions from the photographs and sub-images. Yang et al. [13] proposed a graphical model that displays the correlations between visual features and friends' interactions (i.e. comments) on shared images. Features (e.g. saturation, saturation contrast, bright contrast, cool colour ratio, figure–ground colour difference, figure–ground area difference and background and foreground texture complexity) were utilised as visual characteristics. The model can distinguish between comments that are directly linked to an image's emotional expression and those that are unrelated.

Most cited VSA projects include many handcrafted visual qualities. Although a previous study has demonstrated that all examined elements have a directly influenced the perceived feeling, how to choose the most relevant visual characteristics to the intended task has not yet reached a consensus. Although most previously investigated features are beneficial, recent research on VSA has indicated that employing mid-level representations as a bridge amongst low-level visual features and sentiment orientation is worth investigating. Borth et al. [142] generated a large VSO utilising psychological theories and web mining (SentiBank). 'Beautiful flowers' or 'sad eyes' are examples of ANP. The authors trained a collection of 1.200 visual concept detectors to determine the emotion associated with each picture. ANP detector outputs with a dimension of 1.200 may be used to train a sentiment classifier. They collected adjectives and nouns from YouTube and Flickr video and picture tags. The Plutchik Wheel of Emotion, a well-known psychological model of human emotions, was utilised to find these pictures and videos. The researchers showed a large tagged picture collection of half a million Flickr photographs linked to 1.200 ANPs. The results revealed that SentiBank concepts outperform text-based approaches in tweet sentiment prediction. Zuhe Li et al. [14] developed an approach for completely using the textual sentiment information included in ANPs. They used the image sentiment value as a one-dimensional feature for image sentiment prediction to infer 'the overall sentiment value of an image based on the textual sentiment values of ANPs and the associated replies in the image. The experiments demonstrated that textual SA may improve the accuracy and speed of image SA. Yuan et al. [143] initially used scene-based attributes to generate mid-level features and then built a binary sentiment classifier on top of them. They indicated that adding a step for facial expression recognition increases sentiment prediction when deployed to images featuring faces.

Due to their accessibility and adaptability, VSO and SentiBank are commonly utilised in predicting viewer emotions for photos and animated GIFs [15]. Although most current mid-level representation-based approaches may build a sentiment ontology, they ignore the differences and relationships between ontology concepts. A recent trend in SA and opinion mining for visual information is based on recent achievements in computer vision with DL. Desai et al. [16] intended to solve the issues of VSA by using a DL CNN and affective regions technique to provide intelligible sentiment reports with high accuracy. J. Chen et al. [17] presented a unique active learning architecture that requires minimal labelled training data. Firstly, a new branch called the 'texture module' was added to the CNN. Computing the inner products of feature maps from various convolutional blocks in this branch led to generating the emotional vector, which was then used to distinguish effective pictures. Secondly, a query strategy was constructed using the standard CNN and the texture module classification scores. The samples that produced by the query were then utilised to train the model. Extensive testing on four public affective datasets showed that the method works well for VSA with few labelled training samples. Song et al. [18] introduced sentiment networks with visual attention, a unique architecture that combines visual attention into the well-established CNN sentiment classification framework through end-to-end training. Z. Wu et al. [19] suggested a multitask learning technique for visual attribute identification. The semantic gap between visual characteristics and subjective attributes may be closed by adding sentiment supervision to the attributes. Then, a multi-attention model was used to identify and localise numerous relevant local areas based on

Table 7: Overview of the most significant datasets in VSA

Datasets	Links
International Affective Picture System (IAPS)	`https://csea.phhp.ufl.edu/media.html`
Emotional Category Data from IAPS	`https://link.springer.com/article/10.3758/BF03192732`
Affective Image Classification Dataset	`http://www.imageemotion.org/`
Flickr Sentiments	`http://www.l3s.de/_minack/flickr-sentiment/`
Geneva Affective Picture Database	`https://www.unige.ch/cisa/index.php/downloadfile/view/288/296/`
VSO	`https://visual-sentiment-ontology.appspot.com/`
Emotion6	`http://chenlab.ece.cornell.edu/downloads.html`
Twitter Images	`https://www.cs.rochester.edu/u/qyou/DeepSent/deepsentiment.html`
Cross Sentiment	`http://mm.doshisha.ac.jp/senti/CrossSentiment.html`
T4SA	`http://www.t4sa.it/`
Flickr 8k Dataset	`https://academictorrents.com/details/9dea07ba660a722ae1008c4c8afdd303b6f6e53b`
POM Movie Review Dataset	`http://multicomp.cs.cmu.edu/resources/pom-dataset/`
Flickr Image Dataset for VSO	`https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html`
Twitter Image Dataset for VSO	`https://www.ee.columbia.edu/ln/dvmm/vso/download/twitter_dataset.html`
ICT YouTube Opinion Dataset	`http://multicomp.cs.cmu.edu/resources/youtube-dataset-2/`

expected attributes. The classifier constructed on top of these areas considerably improves visual sentiment prediction. The strategy is better than earlier methods in experiments. Cetinic et al. [20] used CNNs to estimate scores associated with three subjective aspects of human perception: image aesthetic evaluation, mood evoked by the image and image memorability. For each topic, many unique CNN models were trained on various natural image datasets, and the model with the highest performance was selected based on qualitative results and compared to existing subjective artwork ratings. J. Yang et al. [21] described a strategy for locating emotional zones using a deep framework. They used a regular tool to generate N object propositions from a query image and scored them according to abjectness. A pre-trained and fine-tuned CNN model was then used to generate the sentiment score for each proposition. On the basis of both scores, the top K locations were selected from a pool of N applicants. Finally, deep features were taken from the whole image, as well as selected regions, and a sentiment label was projected. Numerous large-scale datasets have demonstrated that the approach can identify emotional local regions and achieve state-of-the-art results. Ou et al. [22] proposed a multilevel context pyramid network to improve the classification performance of VSA. To begin, they utilised Resnet101 to obtain multilevel emotional representation. Subsequently, multiscale adaptive context modules were used to determine the degree of sentiment connection between diverse sections of varying sizes in the image. Finally, many layers of contextual information were combined to create a multi-cue emotional feature for categorisation of image sentiment. Extensive testing on typical visual sentiment datasets showed the efficiency of the strategy. Yadav et al. [23] proposed a residual attention-based DL network for VSA to use CNN to learn the spatial hierarchies of visual features. A residual attention model was employed to focus on the image's critical sentiment-rich local areas. This study made a significant addition by conducting a comprehensive investigation of seven major CNN-based architectures, including VGG-16, VGG-19, Inception-Resnet-V2, Inception-Resnet-V3, ResNet-50, Xception and NASNet. The influence of fine-tuning on these CNN variants was presented in the VSA domain.

3.2.4 VISUAL DATASETS

SA datasets may be collected from various sources. The traditional method of acquiring a set of human-labelled data is to perform large-scale surveys on a massive group of people. Conversely, in SA, a large amount of opinion information can be obtained by leveraging popular social networks (e.g. Instagram, Flickr, Twitter and Facebook) along with webpages devoted solely to gathering business and product reviews. Table 7 provides an overview of the most significant datasets in VSA.

3.3 VISUAL-TEXTUAL SENTIMENT ANALYSIS

People share their ideas on social networking sites, such as Twitter, Facebook and Instagram. These user-generated materials have become more diverse in terms of substance and structure, with users increasingly posting text-embedded pictures called 'image–text postings'. Different from traditional text-only articles, these blogs are more informative because they provide visual materials in addition to text. SA attempts to automatically determine the underlying sentiment of the postings by combining textual and visual materials that may contribute more in understanding user feelings and behaviour. The integration of multiple social contents, their associated characteristics or decisions for performing an analytical task is referred to as multimodal fusion. Numerous definitions of information fusion have been presented in the literature based on [151]. Combining many modalities to achieve good performance in various applications is better than using a single modality. This method has attracted increasing attention from researchers across a range of disciplines due to its potential for an infinite number of applications, including SA, emotion recognition, semantic concept detection, event detection, human tracking, image segmentation and video classification. Information fusion can improve sentiment performance by assuming that the heterogeneity of many information sources allows the cross correction of certain errors, which lead to enhanced results. Utilising many information modalities may provide complementary information and improve the accuracy of the overall decision-making process. To establish the most appropriate fusion approach, the following fundamental questions should be addressed: What is the suitable level to implement the fusion strategy, and how can the information be fused?

For the task of visual–textual sentiment classification, three levels of fusion, namely, early- or feature-level fusion, intermediate- or joint-level fusion and late- or decision-level fusion, can be defined based on the type of available information in a certain field. Table 8 shows an overview of these fusion techniques along with their advantages and disadvantages.

3.3.1 MULTIMODAL FUSION METHODS

This section discusses many fusion strategies that have been used in various visual–textual SA tasks. The fusion approaches are classified as follows: rule-, classification-, attention- and bilinear pooling-based approaches. The optimal strategy for a specific application depends on the nature of the issue, the characteristics of the media and the available parameters.

3.3.1.1) RULE-BASED FUSION METHODS

A number of essential requirements for combining multimodal data are included in rule-based fusion techniques. These methods include linear weighted fusion (sum and product), MAX, MIN, AND, OR, majority voting and custom-defined rules. These concepts were conceptually presented by Kittler et al. [152]. The linear weighted fusion approach is one of the most basic and widely used strategies. It uses sum or product operators to fuse the information received from distinct modalities or decisions derived from a classifier. To combine the data, normalised weights should be allocated to individual modalities. Weight normalisation approaches published in the literature include min–max, decimal scaling, z score, tanh estimators and sigmoid function. Each of these methods has advantages and disadvantages. When the matching scores for the different modalities are easily acquired, the min–max, decimal scaling and z scoring techniques are preferred. However, these strategies are sensitive to outliers. Although the tanh normalisation strategy is robust and efficient, it does need parameter estimation via training. Linear weighted fusion approach is computationally less costly than other methods. However, the weights must be properly normalised for efficient execution. Majority voting fusion is employed based on the decision reached by a majority of the classifiers. Custom-defined rules are application-specific in the sense that they are generated based on information gathered from multiple modalities and the intended ultimate consequence to obtain optimum judgements similar to the work in A. Kumar et al. [153].

3.3.1.2) CLASSIFICATION-BASED FUSION METHODS

This approach uses various classification methods to categorise multimodal data into predefined groups. SVMs, Bayesian inference, Dempster–Shafer theory, dynamic BN (DBN), NNs and ME models are all examples of approaches included in this category. SVM is possibly the most extensively utilised approach for supervised learning on data categorisation jobs. The approach classifies the incoming data vectors into pre-set learnt classes, thereby resolving the pattern classification issue in the context of multimodal fusion. The Bayesian inference fusion approach combines multimodal data according to the probability theory's criteria. This approach integrates features from many modalities or classifiers to produce a joint likelihood inference. The DBN is a network of graphs in which the nodes represent distinct modalities and the connections signify their probabilistic interdependence. The benefit of this network over other approaches is that it allows for easy integration of the temporal dynamics of multimodal data. The second approach that is often employed is NNs. A common NN model comprises input, hidden and output layers. The network

Table 8: Most popular fusion techniques in visual–textual SA

Fusion	Description	Advantages	Disadvantages
Early- or feature-level fusion	It is the process of combining the features extracted from several sources into a single feature vector, which is then fed into an ML algorithm.	• Can utilise the correlation between multiple features from different modalities at an early stage, which helps in better task accomplishment • Requires only one learning phase on the combined feature vector • Does not always need extensive design work • Does not need the training of several models	• Cannot capture the complementary nature of the modalities • Generates high-dimensional feature vector that may even contain redundancies • Requires all features to have the same format before initialising the fusion process • Cannot capture the time-synchronicity of several sources due to the fact that the characteristics from separate but tightly connected modalities might be retrieved at different times • Increase in the number of modalities make it difficult to learn the cross-correlation amongst the heterogeneous features
Intermediate- or joint-level fusion	It is the process of combining learned feature representation from intermediate layers of NNs with features from other modalities as input to a final model. Most of the intermediate fusion models adopt a shared representation layer to merge units with connections from multiple modality-specific paths.	• Can simulate interactions between features from several modalities • Can learn more enhanced feature representation from each modality • Can combine input at various abstraction levels • Does not require the training of several models	• Might result in overfitting, in which the network fails to simulate the relationship between both modalities; therefore, a careful design is required • Cannot perform SA effectively when a portion of multimodal information is missing
Late- or decision-level fusion	It is the process of integrating the results of numerous sentiment classifiers that have been trained independently for each modality.	• Can predict even when not all modalities exist • Having a large quantity of training data is not required • No need to convert data to the same format • More flexibility as using the most suitable methods for analysing each single modality • Scalable (i.e. graceful upgradation or degradation) in terms of the modalities used in the fusion process	• Cannot capture the relationship between several modalities • Ignores the modalities' low-level interaction • Ensemble all classifiers effectively is difficult • Probability of missing local interactions between modalities

may be fed with data in the form of multimodal characteristics or classification choices from several classifiers. Results contain contains a fusion of the data being considered. The activation functions required to generate the predicted output are supplied by the hidden layer of neurons, and the number of neurons and hidden layers is selected to provide the required level of accuracy. The technique is most often used for decision-level fusion.

3.3.1.3) ATTENTION-BASED FUSION

The fusion process mostly uses the attention mechanism like in K. Zhang et al. [154] work, which refers to the weighted sum of a collection of vectors with scalar weights dynamically created at each step by a small 'attention' model. Numerous glimpses (output heads) are often utilised to create numerous sets of dynamic weights for summing, which might retain additional information by combining the results from each glimpse.

When attention mechanisms are applied to an image similar to the work in K. Zhang et al. [155], the feature vectors associated with distinct areas are weighted differently to create an attended image vector, which is used to correlate the emotional image regions with the corresponding text description.

In contrast to image attention mechanisms, M. Cao et al. [156] applied the co-attention mechanisms which utilise a symmetric attention framework to generate attended image and text feature vectors, whereas J. Xu et al. [157] used the dual attention network (DAN) which is similar to the parallel co-attention in predicting the attention distribution for pictures and languages concurrently. Such attention models are based on the characteristics and memory vectors of modalities. Different from co-attention, memory vectors may be repeatedly modified at every reasoning level by utilizing repetitive DAN structures.

Conversely, J. Xu, et al. [158] presented an alternating co-attention mechanism that generates an attended image vector based on linguistic characteristics, followed by an attended language vector based on the attended image vector.

Multi-head attention was also proposed by X. Yang et al. [159] and S. Zhang et al. [160] to define the correlation between the image and text contents. Furthermore, gated multimodal fusion is considered a different type of attention because it uses gating to assign weights to visual and textual elements. The weighted sum of the visual and textual feature vectors may be computed based on the dimension-specific scalar weights created automatically using a gated process. Then, multimodal sentiment classification may be performed using these representations like in F. Huang et al. [161] and J. Arevalo et al. [162] works.

3.3.1.4) BILINEAR POOLING-BASED FUSION

Bilinear pooling, also called order pooling [163], is a technique for fusing visual and textual feature vectors into a shared representation space by calculating their outer product, allowing for multiplicative interactions between all items in both vectors. Bilinear pooling generates a (n2) dimensional representation by linearising the matrix generated by the outer product into a vector, which is more expressive than simple vector combination operations (assuming each vector has n elements), such as a weighted sum, element-wise multiplication or concatenation, which result in n or 2n-dimensional representations. Bilinear representations are usually linearly converted to output vectors using a two-dimensional weight matrix, which is analogous to fusing two input feature vectors using only a three-dimensional tensor function. When computing an outer product, every feature vector may be expanded by one to retain single-modal input features [164]. Conversely, bilinear pooling needs the decomposition of weight tensors to enable the associated model to be trained appropriately and effectively due to its enormous dimensionality (usually on the scale of hundreds of thousands to millions of dimensions) [165]. Bilinear pooling can be combined with attention mechanisms [166, 167, 168] to obtain optimum correlation between visual and textual contents.

Although SA upon single modality data has seen significant success in recent years, it cannot successfully manage the variety of information in social media data. As a result, multimodal SA arose at the right time, which has gained plenty of attention in recent years. Zhou et al. [166] presented a cross-modality hierarchical interaction model for SA. The method handles noise and joint comprehension concerns by extracting semantic (bilinear attention mechanism) and sentiment interactions (multimodal CNN) between picture and text in a hierarchical approach. A hierarchical attention technique Was initially used to collect semantic connection and filter information in one mode using another. Then, a multimodal CNN was used to fully leverage cross-modal sentiment connection, resulting in superior visual—textual representation. A transfer learning strategy was further developed to reduce the influence of noise in actual social data.

N. Xu et al. [169] characterised the relationship between text and picture using a co-memory attentional method. Despite the fact that the reciprocal influence of text and picture was considered, a coarse-grained attention mechanism was applied, which may not be able to retrieve sufficient information. Conversely, the co-memory network merely uses a weighted textual/visual vector as a guide to learn attention weights on visual/textual representation. It may be thought of as a coarse-grained attention mechanism that can lead to information loss because attending several items with a single attention vector might obscure the characteristics of each attended content.

Jiang et al. [170] suggested a fusion extraction network model for multimodal SA. To begin, the model employs an interactive information fusion technique to learn the visual-specific textual and textual-specific visual representations dynamically. Then, for the particular textual and visual representations, an information extraction technique is utilised to extract meaningful information and filter out unnecessary components.

Zhao et al. [24] presented a novel image–text consistency-driven multimodal SA technique to investigate the connection between the picture and the text, followed by a method for multimodal adaptive SA. The standard SentiBank technique was utilised to describe the visual cues by extracting the mid-level visual features with the corporation of additional characteristics, such as textual, low-level visual and social characteristics to construct an ML SA approach utilising the VSO as a benchmark dataset.

Guo et al. [171] proposed an end-to-end news sentiment recognition method using Layout-Driven Multimodal Attention Network (LD-MAN). Instead of modelling text and pictures separately, LD-MAN aligns images with text using online news layout. LD-MAN uses a set of distance-based coefficients to represent picture locations and quantify image–text contextual relationships. LD-MAN then learns the articles' emotive representations from aligned text and graphics using multimodal attention.

Ke Zhang et al. [155] suggested a unique cross-modal semantic content correlation strategy based on deep matching and hierarchical networks. The proposed joint attention network can learn the content association between an image and its caption, which is exported as a pair of images and texts. The caption was processed by a class-aware sentence representation (CASR) network equipped with a class dictionary, and a fully connected layer concatenated the CASR outputs into a class-aware vector. The class-aware distributed vector was finally queried with the image–text pair as input into an inner-class dependency LSTM to extract cross-modal nonlinear correlations for sentiment prediction.

F. Huang et al. [161] suggested a unique method called attention-based modality-gated networks (AMGN) to utilise the relationship between both modalities and extract the discriminative features for multimodal SA. To learn attended visual aspects for each word, a visual semantic attention model was presented. A modality gated LSTM Was suggested to learn multimodal features by adaptively picking the modality that yields higher sentiment information. An automated semantic self-attention model was then suggested to put more attention on the discriminative features for sentiment classification.

You et al. [172] suggested a model for cross-modality consistent regression (CCR) that incorporates visual–textual SA methods. Then, a CNN was used to extract the visual characteristics, and the textual features were learned using a distributed paragraph vector model trained on the titles and descriptions of images. To conduct integrated visual–textual SA, several fusion approaches, such as early fusion, late fusion and CCR, were used. By enforcing consistent constraints across related modalities, the CCR model was trained to develop the final sentiment classifier.

Arevalo et al. [162] introduced a unique multimodal learning model based on gated NNs. The gated multimodal units (GMU) model was meant to be utilised as an intrinsic unit in a NN architecture with the task of constructing an intermediate representation from a collection of input from several modalities. By using multiplicative gates, the GMU learns to determine how modalities affect the activation of the unit.

Duong et al. [173] developed basic approaches (i.e. common feature space Fusion using an auxiliary learning task and joint fusion using a pooling layer) for classifying social media material that incorporate information from several modalities. Baecchi et al. [174] investigated the application of multimodal feature learning techniques, such as SG and denoising autoencoders, for SA of micro-blogging material, such as short Twitter messages, which include text and perhaps an image. A unique architecture that combines these NNs was presented, and its efficiency and classification accuracy on numerous typical Twitter datasets was demonstrated.

Peng et al. [175] presented a cross-modal complementary network (CMCN) with hierarchical fusion for MSC. The CMCN is built as a hierarchical structure with three major modules: feature extraction from texts and pictures, feature attention from an image–text correlation generator and cross-modal hierarchical fusion. In this manner, a CMCN may assist in limiting the risk of incorporating unrelated modal features, and it outperforms existing approaches.

Ji et al. [176] presented a unique bilayer multimodal hypergraph learning (bi-MHG) for robust sentiment prediction of multimodal tweets. A two-layer framework was developed for the suggested bi-MHG model (tweet-level hypergraph and feature-level hypergraph). In a bilayer learning system, multimodal characteristics are shared between the two layers. Bi-MHG specifically designs the importance of modality rather than intuitively weighting multimodal information as in conventional multimodal hypergraph learning. Finally, layered alternating optimisation was presented for parameter learning.

J. Xu et al. [177] suggested a unique hierarchical deep fusion model to study cross-modal connections among pictures, words and social ties for more effective SA. Specifically, they used three-level hierarchical LSTMs to learn inter-modal correlations between picture and text. A weighed relation network with each node embedded in a distributed vector was used to efficiently use connection information. The obtained image–text features and node embeddings were fused using a multilayer perceptron (MLP) to further capture nonlinear cross-modal correlations.

Kang Zhang et al. [154] developed a novel multimodal sentiment model that can remove the noise from textual input whilst extracting important visual characteristics. Then, they used the attention mechanism for feature fusion where the textual and visual contents learn the intrinsic characteristics from each other by symmetry. Fusion characteristics were then used to classify the sentiments.

J. Xu et al. [157] suggested an attention-based heterogeneous relational model to incorporate rich social information into multimodal SA to increase performance. A progressive dual attention module was designed to record image–text interactions and subsequently train the combined image–text representation from the standpoint of content information. Here, a channel attention schema was provided to emphasise semantically-rich picture channels, and a region attention schema was presented to identify emotional areas based on the attended channels. The researchers next built a heterogeneous relation network and adapted the graph convolutional network to integrate material information from social contexts as supplements to create improved representations of social pictures.

T. Zhu et al. [168] introduced a novel image–text interaction network to examine the association between expressive picture areas and text for multimodal SA. Specifically, a cross-modal alignment module was used to capture area word correspondence, which was then fused by an adaptive cross-modal gating module. They also incorporated individual-modal contextual feature representations to provide a more accurate prediction. Tashu et al. [178] demonstrated a multimodal emotion detection architecture that leverages feature-level (sequential co-attention) and modality attention to categorise emotion in art. The proposed architecture allows the model to learn informative and refined representations for feature extraction and modality fusion. The proposed approach performed well on the WikiArt emotion dataset.

Ortis et al. [179] suggested extracting and applying an Objective Text description of pictures instead of the traditional Subjective Text given by users to estimate sentiment associated with social images. Objective text is extracted from photos using modern DL architectures to categorize objects and scenes, and conduct image captioning. The objective text characteristics are merged with visual information in a canonical correlation analysis embedding space. Then, SVM was used to infer the sentiment polarity. The research demonstrated that using text retrieved from photos instead of text given by users enhances classifier performance.

Cai et al. [180] proposed a sentiment prediction system based on CNN for multimedia data (tweets containing text and images). They employed two independent CNNs to learn linguistic and visual properties and then combined their outputs into another CNN to completely investigate the text–image relationships. You et al. [167] offered a novel approach for robust SA that combines text and visual data. Different from earlier research, the authors claimed that visual and textual information should be addressed structurally. To perform a correct analysis, they initially constructed a semantic tree structure using sentence parsing. The system then learned a strong visual–textual semantic representation by combining an attention mechanism with LSTM and an auxiliary semantic learning task.

X. Yang et al. [181] developed an innovative multimodal emotion analysis model based on the multi-view attentional network for deep semantic properties of image and text. The framework consists of three steps, namely, features mapping, interactive learning and feature fusion, which make use of stacking-pooling and multilayer perceptrons to profoundly fuse the various features. As part of this study, TumEmo image–text emotion dataset was created.

Yadav and Vishwakarma [182] presented a deep multilevel attentive network to analyse visual and textual correlations. A bi-attentive visual map was created using spatial and channel dimensions to boost CNN's representation capacity. Then, the association between the bi-attentive visual features and word semantics was established by applying semantic attention. Finally, self-attention was used to automatically harvest sentiment-rich multimodal characteristics. Miao et al. [183] determined the essential areas of an image based on the its word representation. The picture portions were considered as memory cells, and the attention process was used to retrieve them. Then, CNN was used to blend visual and textual information more efficiently using the IMDB dataset.

F. Huang et al. [184] presented a new visual–textual SA model called deep multimodal attentive fusion, which uses visual and semantic internal correlations to analyse sentiments. To acquire efficient emotion classifiers for the visual and textual modality, two distinct unimodal attention models were presented. Then, for joint sentiment classification, an intermediate fusion-based multimodal attention model was proposed. Finally, late fusion was used to merge the three attention models.

X. Zhu et al. [185] suggested that the influence of visual and textual information to SA should be differentiated. By including a cross-modal attention mechanism and semantic embedding knowledge through a bidirectional RNN, the model develops a strong joint visual–textual representation, making it better than the current state-of-the-art methods.

MultiSentiNet is a deep semantic network proposed for multimodal SA by N. Xu and Mao [186]. The authors used salient object and scene detectors to extract deep semantic information from photos. Then, a visual feature-guided attention LSTM model was proposed to extract important words for comprehending the emotion of a whole tweet and integrating the representation of those informative words with visual semantic characteristics, objects and scenes. Experiments showed the efficiency of the MultiSentiNet model.

X, Yang et al. [159] developed a multichannel graph NNs with sentiment awareness for image–text sentiment detection. To capture hidden representations, they initially encoded multiple modalities. Then, multichannel graph NNs were used to learn multimodal representations using global dataset properties. Finally, multimodal in-depth fusion was used to estimate the emotion of image–text pairings.

The various Syncretic co-attention network was presented by M. Cao et al. [156] to uncover multilevel corresponding interactions between multimodal data and incorporate the unique information of each modality for unified complementary sentiment classification. A multilevel co-attention module was constructed to study localised correspondences between picture regions and text words, as well as holistic correspondences across global visual data and context-based textual meaning. Then, all single-modal characteristics may be combined from multiple levels. The suggested VSCN additionally incorporates unique information of each modality concurrently and combines them into an end-to-end system for SA.

Ye et al. [165] developed visual–textual sentiment based on product reviews. Their study has two significant contributions. Firstly, a new dataset called Product Reviews was developed instead of scraping data from Flickr or Twitter with positive and negative labels. Secondly, a deep Tucker fusion approach was proposed for visual–textual SA using Tucker decomposition and bilinear pooling operation to integrate deep visual and textual representations. J. Xu et al. [158] introduced a unique bidirectional multilevel attention model to jointly classify visual and textual sentiments by using complimentary and complete information from image and text modalities.

A. Kumar et al. [153] proposed a hybrid DL model for fine-grained sentiment prediction in multimodal data using DL networks and ML to handle two distinct systems (i.e. textual and visual) and their combination inside online material utilising decision-level fusion. The proposed contextual ConvNet–SVMBoVW model was trained using comments and postings (text, picture and infographic) generated using the #CWC2019 hashtag on Instagram and Twitter. The accuracy of the proposed model is regarded superior than that of the text and picture modules.

Kumar and Garg [187] suggested a multimodal SA approach to analyse textual–visual data (or infographic and typographic). A CNN was used to score the image sentiment utilising SentiBank and SentiStrength (R-CNN), and an innovative hybrid (lexicon and ML) method was used to score text sentiment. After separating text from images using optical character recognition, multimodal sentiment scoring was performed by combining the sentiment ratings from images and texts. High accuracy was achieved in the suggested model's performance using random multimodal tweet dataset.

Liao et al. [188] presented an image–text interactive graph NN for SA. The graph's node features started by text and picture features and were updated by the graph's attention mechanism. Finally, it would be merged with an image–text aggregation layer to achieve sentiment classification. Yu et al. [189] has suggested a CNN-based system for visual–textual SA. For Chinese microblogging, they built CNN models that combined picture and text characteristics using written and visual information.

Y. Zhang et al. [190] proposed a quantum-inspired multimodal SA (QMSA) system. The system contains a quantum-inspired multimodal representation model. This approach was used to address the semantic gap and represent the correlations between multiple modalities using a density matrix, as well as a multimodal decision fusion technique inspired by quantum interference.

Liu et al. [191] suggested a useful method for the SA of GIF videos with written descriptions that combines visual and linguistic information. By using a sentiment classifier, each GIF video's visual characteristics were retrieved and categorised. The sentiment probability obtained from the visual classifier was further transformed into a sentiment score by developing a mapping function. Simultaneously, the SentiWordNet3.0 model was used to assign a sentiment score to the extracted sets of significant sentiment words from the attached textual annotations. Finally, a sentiment score function comprising visual and textual components was built with a remarkable difference threshold to further improve the fused sentiment score.

Qian et al. [192] used deep CNN trained on Twitter data to provide a unique approach for extracting features from Twitter photos and the accompanying labels or tweets. They fine-tuned AlexNet to extract the visual features, along

Table 9: Most popular datasets used in visual–textual SA

Dataset	Description and Link
Photo Tweet Sentiment Benchmark	A small dataset of 603 picture tweets (with textual data) encompassing a broad variety of subjects was initially provided by D. Borth et al. [142], which is then labelled by Amazon Mechanical Turk (AMT) workers, yielding 470 positive and 133 negative categories. `https://www.ee.columbia.edu/ln/dvmm/vso/download/twitter_dataset.html`
VSO: Image Dataset	The database consists of two datasets: a collection of Flickr photos with Creative Commons licenses that were used to train/test 1,200 ANP detectors in SentiBank and a set of images connected with the entire VSO, which includes 3,244 ANPs [142]. `https://www.dropbox.com/sh/xs049whsb3wt3c8/AAB5TSbydN-6yGqYnHo1XHgCa?dl=0`
Multi-view SA Dataset (MVSA)	An MVSA for visual–textual SA was recently provided by T. Niu et al. [195]. Only text–image tweets with accessible photos were saved for annotation. A total of 4,869 texts contain three sentiments (positive, negative and neutral). `https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/`
T4SA	A huge dataset of over 3 million tweets provided by L. Vadicamo et al. [196] contains text and images that has been categorised by the sentiment polarity of the text (negative = 0, neutral = 1, positive = 2). `http://www.t4sa.it/`

with the use of affective space of English ideas, which were employed to extract the text features. Finally, a unique sentiment score was suggested to integrate the picture and text predictions. The assessment was based on a Twitter dataset that included photos, labels and tweet messages.

S. Zhang et al. [160] proposed a hybrid fusion network to collect inter- and intra-modal characteristics. A multi-head visual attention was presented to derive correct meaning and sentimental insights from text contents, guided by visual cues. In the decision fusion step, several baseline classifiers were trained to learn distinct discriminative knowledge from numerous modal representations. The final decision was made by combining decision supports from the baseline classifiers.

Truong et al. [193] proposed a visual–textual attention network based on the observation that images can play a supporting role for the text. The proposed model depends on visual data to highlight the important texts of the whole document. P. Kumar et al. [194] proposed a visual–textual emotion classification model using DL based on fusion strategies. Firstly, they performed a joint fusion to combine the visual and textual data. Secondly, a decision fusion model was created to combine the results of the textual, visual and joint fusion outputs. A new emotional dataset was generated using the B-T4SA dataset. Table 10 summarises the literature done for visual–textual SA.

3.3.2 VISUAL–TEXTUAL DATASETS

With a large number of social media networks (e.g. Twitter, Flicker, Facebook and Instagram) where people can share their opinions and sentiments on various topics, a large dataset for visual–textual contents may be collected. However, collecting such a large number of datasets requires significant effort for annotation. Some benchmark datasets available for visual–textual SA are presented in Table 9.

3.3.3 EVALUATION MEASURES

Various measures are used to evaluate and compare the model's performance and efficiency. Confusion matrices (also known as error matrices or truth tables) are a well-known method for evaluating and summarizing the performance of a classification model. This matrix displays the number of properly and wrongly categorized samples identified by a classifier. The abbreviations TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative) in the confusion matrix correspond to the following:

- TP: Number of instances where the expected class label is positive while the actual class label is correct.
- FP: Number of instances where the expected class label is positive and the actual class label is incorrect.
- FN: Number of instances where the expected class label is negative and the actual class label is incorrect.
- TN: Number of samples where the expected class label is negative and the actual class label is correct.

Accuracy, Precision, Recall, and F1-score are regularly used performance assessment metrics by many researchers, that are calculated according to the confusion matrix. Accuracy is the ratio of accurately predicted instances to the total number of examples. Its value is determined using equation 1.

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

Precision refers to the total number of correctly expected class labels for each class. The precision value is determined using equation 2.

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

Recall value is the weighted average of properly categorized labels for each class. This value is determined using equation 3.

$$Recall = \frac{TP}{TP + TN} \tag{3}$$

F1-score is used to combine precision and recall values in a single measurement. This number ranges from 0 to 1, and if the classifier correctly classifies all samples, it delivers a value of 1, indicating a high level of classification success. The F1-score value is determined using equation 4.

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

Accuracy is often the most used statistic for measuring algorithm performance. However, other measures (accuracy, recall, and F1-score) were utilized by researchers to evaluate their investigations. For instance, J. Xu et al. [158] and T. Liu et al. [191] evaluated their models using accuracy, precision, recall, and F1-score. In contrast, A. Kumar and G. Garg [187] evaluated the efficacy of their model using an accuracy metric, while T. Zhu et al. [168] employed both accuracy and F1-score for performance assessment.

4 CHALLENGES

Multimodal SA is the extraction of associated sentiments and emotions from the multimodal inputs. Nowadays, most of the posts or blogs on social media are multimodal (i.e. more than one modality is associated with the data), mostly a combination of textual and visual elements. Although plenty advancements have been achieved in this area, many issues and challenges still need to be addressed to improve its performance. Some of the challenges in multimodal SA are as follows.

- Heterogeneity in Feature Space: Different from traditional single-modality SA, multimodal SA includes a varied collection of manifestation patterns. Visual material and textual description are diverse in feature spaces. As a consequence, SA should be effective in bridging the gap between diverse modalities [181, 182, 184].
- Extraction of the Sentiment Intensity: Different semantic information may be covered by the visual content and word description that comes in various forms, such as expressions, emoticons and gestures. The present computational approaches are primarily concerned with extracting syntactic information from user content; however, they fail in capturing the sentiment information, which is critical in the SA process [190].
- High-dimensional Feature Space: As multiple modalities are involved; each modality has a distinguishing property that aids in characterising it. As a result, the same strategies for feature extraction cannot be used to all modalities (i.e. textual and visual). To categorise sentiments in visual–textual SA, feature vectors for each modality are created individually and it is then combined, fusing all of these feature vectors into a single global vector is a time-consuming procedure, which may lead to high-dimensional feature vector that must be solved [24, 177].
- Cross-modal Relationships: For SA, a huge amount of data is accessible through the social platforms using different formats. Users express their opinion/sentiment in several modes concurrently because they aim to transmit the same polarity of feeling across various modes. As a result, the relationship between visual content and its textual semantics should be considered to achieve better sentiment classification [155, 172, 175].
- Incomplete Modality: It is not uncommon for one modality to be absent from multimodal data. Many users will often submit tweets without including photos, whereas photographers will frequently share photos without including a written description. Handling incomplete multimodal data for sentiment classification is another problem [173].

- Domain-independent Sentiment Polarity: Most present research attempts to predict sentiment polarity using domain knowledge. The system cannot infer sentiments from microblogs if it was trained on a plain text content from documents. This characteristic prevents it from adapting to new domains [197].
- Integrating Data from Several Sources: The fusion approach should be used to integrate the information gathered from all modalities. Determining which modality should contribute the most to the fusion by assigning weightages to each modality is important to examine carefully throughout the fusion process. In feature-level fusion, the feature vectors with the largest weighted modality will be the dominant element in the fusion process. Moreover, the sentiment score of the modality with the greatest weight should be included in decision-level fusion [161, 155, 172].
- Dataset Scarcity: There exists a shortage on the available datasets, especially in multimodal SA using visual and textual content, and most of the existing datasets suffer from many problems. One of these problems is that imbalance in the class distribution will affect the performance of ML-based sentiment prediction. Another issue is the subjectivity of human emotion, which causes label unreliability. Even manually annotated sentiment labels are not 100% trustworthy [142, 193].
- Noise Presented with Additional Modalities: Images and words may be quite semantically expressive. In most circumstances, the usable semantic segment for SA is only a portion of the picture or text. The remainder is redundant and may lead to sentiment misinterpretation errors. The rising volume of data in visual–textual SA exacerbates this difficulty. Misunderstanding one modality might lead to poor results even when the other is correctly assessed. A unique method is required to combat multimodal noise [166].

5 APPLICATIONS OF SENTIMENT ANALYSIS

The increasing availability of emotional data from numerous forums, blogs and social networks has raised academic and industry awareness of SA. SA may help organisations understand people's attitudes and preferences based on prior behaviour. Thus, it may enable them to personalise their goods and services to their needs. The various application fields of SA are as follows.

- Predictions for the Box Office. With the fast growth of social media, many online film reviews are accessible in text and video format, which enables forecasting the box office success of films. A simple sentiment-aware autoregressive model was presented by P. Nagamma et al. [198], which used TF-IDF values as features and Fuzzy Clustering. An SVM classifier was also developed for forecasting box office revenue trends based on review sentiment. H. Timani et al. [199], three sentiment indexes were created using YouTube movie trailer comments to quantify movie reviews' sentiment. These indexes were then used to predict a movie's box office collection. The comments on the trailers assisted distributors and filmmakers in estimating the movie's reaction rate.
- Forecasting the Stock Market. Despite the advancement of the computer world, the unpredictable characteristic of the stock market makes forecasting one of the most challenging undertakings. SA may assist in the development of efficient methods. In A. H. Moghaddam et al. [200], an ANNs, which were trained via backpropagation, were used to forecast the NASDAQ stock market index.
- Business Intelligence. Many firms are now using SA to aid in decision making and business improvement. In P. P. Rokade and D. Aruna Kumari [201], a revolutionary approach was proposed to business analytics in modern enterprises.
- Summarisation of Television Programmes and Newscasts. A system that uses multimodal SA to automatically assess broadcast video news and construct summaries of television shows was proposed by J. G. Ellis et al. [202]. They described a method for analysing and creating person-specific fragments from news video, delivering 929 sentence-length videos annotated using Amazon Mechanical Turk.
- Recommender Systems. Numerous applications provide suggestions based on users' prior experience. For instance, in the retail business, consumers who seek for a certain product may obtain suggestions for future endeavours. In X. L. Zheng et al. [203], a hybrid technique that offers correct recommendations was presented.
- Prediction of Political Trends. The growing popularity of social media platforms improves the potential of forecasting the result of an election. For example, P. Chauhan et al. [204] wrote a survey article that discusses the assessment of SA methodologies and attempts to demonstrate the researchers' contribution to forecasting election outcomes using social media information. Moreover, this article makes some recommendations for future research on election prediction using social media information.
- Healthcare. It is the most frequently utilised domain, which is used to assess patient evaluations regarding their health that are shared on different types of social media sites (e.g. Twitter, Facebook and Instagram). SA

model was developed by F. J. Ramirez-Tinoco et al. [205] to obtain sentiment and emotional information that may aid healthcare practitioners in comprehending their patients' emotions and concerns by taking appropriate remedial action.

6 CONCLUSION

This paper provides an overview of visual–textual SA and their related techniques by reviewing the current works to provide researchers with a complete understanding of the methodology and resources available for visual and textual SA. It also covers the process of categorising and summarising the most widely used SA methodologies and their benefits and limitations, including important procedures, such as pre-processing, feature extraction, data fusion strategies and the most widely used sentimental datasets. It also examines some of the field's most pressing challenges and applications.

Our investigation of this field reveals its potential for using complementary information channels for SA, which is often better than unimodal approaches. It can also improve other tools that instantly benefit from unimodal SA, such as entity recognition and subjectivity analysis. We hope that our review will encourage further cross-disciplinary collaboration in this emerging industry. Future research may include further work on assessing sentiments from other domains. A number of case studies may be explored to determine the usefulness of various methodologies in investigating sentiments.

Table 10: Summary of the literature for visual-textual SA

Author	Features	Classification Approach	Fusion	Dataset	Performance Metric
Zhou et al. [166]	Visual: VGG19 Textual: GloVe	MLP	Joint fusion	Getty Image VSO_VT	Accuracy: 0.936 Accuracy: 0.872
N. Xu et al. [169]	Visual: pre-trained CNN Textual: GloVe	Soft-max Classifier	Early Fusion	MVSA-single MVSA-multiple	Accuracy: 70.51% Accuracy: 68.92%
Jiang et al. [170]	Visual: ResNet152 Textual: GloVe and Bert	Soft-max Classifier	Joint Fusion	MVSA-single MVSA-multiple	FENet-GloVe 72.54% and BERT 74.21% FENet–GloVe 70.57% and BERT 71.46%
Zhao et al. [24]	Visual: Low-, Mid- and High-level Feature Textual: Word2vec Social: Lifespan, Emotion and Reach	SVM	Joint Fusion	VSO	Accuracy: 87%
Guo et al. [171]	Visual: ResNet152 Pre-trained on Place365 and ImageNet Textual: GloVe	Fully connected NN	Joint Fusion	RON DMON Collected / multimodal online news datasets	Avg. Accuracy: 53.51% for RON Avg. Accuracy: 80.81% for DMON
Ke Zhang et al. [155]	Visual: VGG-19 Textual: GloVe Class Dictionary	IDLSTM: Inner-Class Dependency	Early Fusion	Flicker Getty-image Twitter	Accuracy: 0.842 Accuracy: 0.806 Accuracy: 0.863
F. Huang et al. [161]	Visual: VGG-19. Textual: GloVe	LSTM. Soft-max Classifier	Joint Fusion (AMGN)	Twitter-W Getty Images -W Flickr-W Flickr-M	Accuracy: 79% Accuracy: 88% Accuracy: 87% Accuracy: 89%
You et al. [172]	Visual: Pre-trained CNN Textual: Word2vec was initially employed and then taking the average	Soft-max Classifier	Joint Fusion (CCR)	Getty Images Twitter AMT-Twitter	Accuracy: 0.800 Accuracy: 0.809 Accuracy: 0.769

Arevalo et al. [162]	Visual: VGG, End2End CNN Textual: n-gram, word2vec, RNN_w2v, RNN_end2end	Logistic regression and MaxoutMLP for early and joint fusion. Average prediction for late fusion	Early fusion Joint fusion (GMU) Late fusion	Multimodal IMDB	F-score: 0.600, 0.600 for early fusion using CONCAT and Linear Sum. F-score: 0.604 for late fusion. F-score: 0.617 for GMU model.
Duong et al. [173]	Visual: Inception Textual: GloVe	NNs	Joint fusion and Common Feature Space Fusion	Flicker Reddit Collect/ Emotion dataset	Accuracy: 93.01%, 93.44% for Flicker using joint and CFS fusions Accuracy: 86.29%, 86.92% for Reddit using joint and CFS fusion
Baecchi et al. [174]	Visual: Denoising autoencoder Textual: Word2vec (CBOW)	Logistic regression	Early fusion	SentiBank Twitter Dataset	Accuracy: 79%
Peng et al. [175]	Visual: VGG Textual: Bert	Soft-max Classifier	Early fusion	MVSA-single MVSA-multiple Multi-ZOL	Accuracy: 73.61% Accuracy: 70.45% Accuracy: 74.28%
Ji et al. [176]	Textual: bag of textual words Visual: BOVW using SentiBank Emoticon: bag of emoticon words	Bi-Layer Multimodal Hypergraph Learning (Bi-MHG)	Early fusion	Cross-modality Chinese dataset/ Sina Weibo platform	Accuracy: 90.0%
J. Xu et al. [177]	Visual: VGG19 Textual: GloVe Network: network embedding method	MLP	Joint fusion	Flicker Twitter Flicker-Ml	Accuracy: 0.859 Accuracy: 0.767 Accuracy: 0.881
Kang Zhang et al. [154]	Textual: Denoising Autoencoder Visual: Attention-based Variational Autoencoder	Soft-max Classifier	Early fusion	MVSA-single MVSA-multiple	Accuracy: 0.7144 Accuracy: 0.6962
J. Xu et al. [157]	Visual: VGG19 Textual: GloVe Social relation features extracted using GCN	MLP	Joint fusion	Flickr Getty Image	Accuracy: 0.871 Accuracy: 0.878
Tong Zhu et al. [168]	Visual: Faster R-CNN which is pretrained on Visual genomes dataset using ResNet-101 as backbone Textual: BERT-Base	Soft-max Classifier	Joint fusion	MVSA-single MVSA-multiple	Accuracy: 0.7519 Accuracy: 0.7352
Tashu et al. [178]	Visual: ResNet50 Textual: GloVe and Emotion Category extracted using three layer FFNN	Soft-max Classifier	Joint fusion	Wiki-Art Emotions	Accuracy: 0.773

Ortis et al. [179]	Visual: RGB histogram, GIST descriptor, BOW image, SentiBank 1200, GoogLeNet, DeepSentiBank, and Places205 Textual: BOW, SentiwordNet, and word dictionary	Linear SVM	Joint fusion	Flickr	73.96 ±0.39% &72.66 ±0.70% represents the average and STD for the full features & the truncated features
Guoyong Cai and Binbin Xia [180]	Visual: CNN Textual: Word2vec	Soft-max Classifier Logistic regression	Joint fusion	Twitter (TD1) Twitter (TD2)	Accuracy: 0.78 Accuracy: 0.796
You et al. [167]	Visual: VGG19. Textual: One-Hot Encoding, GloVe	Soft-max Classifier	Joint fusion	Getty Twitter Twitter AMT VSO-VT	Accuracy: 0.902 Accuracy: 0.964 Accuracy: 0.904 Accuracy: 0.833
X. Yang et al. [181]	Visual: VGG-Object and VGG-Place Textual: GloVe	Multilayer Perceptron and a stacking-pooling module	Joint fusion	MVSA-single MVSA-multiple Tum-Emo	Accuracy: 0.7298 Accuracy: 0.7236 Accuracy: 0.6646
Yadav and Vishwakarma [182]	Visual: Inception V3 Textual: GloVe	Soft-max Classifier	Joint fusion	MVSA-single MVSA-multiple Flickr Getty Image	Accuracy: 79.47% Accuracy: 77.89% Accuracy: 89.30% Accuracy: 92.65%
Miao et al. [183]	Visual: VGG19 Textual: GloVe	Dense Layer	Joint fusion	MM_IMDB	Accuracy: 0.642
F. Huang et al. [184]	Visual: VGG19 Textual: GloVe	Fully connected (FC) layer for joint fusion then a custom defined rules are used to fuse the visual, textual and joint fusion parts.	Combining joint fusion and late fusion	Twitter-W Getty Images-W Flickr-W Flickr-M	Accuracy: 0.763 Accuracy: 0.869 Accuracy: 0.859 Accuracy: 0.880
X. Zhu et al. [185]	Visual: Inception-V3 Textual: GloVe	Two-layer Perceptron	Joint fusion	Getty Images VSO	Accuracy: 0.913 Accuracy: 0.851
Xu and Mao [186]	Visual: VGG19-ImageNet & VGG19-Place Textual: GloVe	Soft-max Classifier	Joint fusion	MVSA-single MVSA-multiple	Accuracy: 69.84% Accuracy: 68.86%
X. Yang et al. [159]	Visual: YOLOv3, VGG-Place, and ResNet used for object and scene memory banks Textual: GloVe	Soft-max Classifier	Joint fusion	MVSA-single MVSA-multiple Tum-Emo	Accuracy: 0.7377 Accuracy: 0.7249 Accuracy: 0.6672

M. Cao et al. [156]	Visual: local (RPN) and Global (ResNet101) Textual: GloVe (word embedding) and LSTM for sentence representation	Multilayer perceptron for fusion and Soft-max Layer for classification.	Joint fusion	Twitter-W Getty Images-W. Flickr-W. VSO- Strong.	Accuracy: 0.824 Accuracy: 0.856 Accuracy: 0.841 Accuracy: 0.905
Ye et al. [165]	Visual: ResNet101 Textual: nn.embedding layer in Pytorch	Soft-max Classifier	Joint fusion	PR-150K MVSO VSO	Accuracy: 0.733 Accuracy: 0.801 Accuracy: 0.856
J. Xu et al. [158]	Visual: Local (Faster R-CNN) and Global (VGG19) Textual: GloVe	Soft-max Classifier	Joint fusion	Flickr-W Flickr-ML Getty Images-W Flickr-IML	Accuracy: 0.849 Accuracy: 0.878 Accuracy: 0.865 Accuracy: 0.831
A. Kumar et al. [153]	Visual: LBP descriptor used for extracting the visual features to construct the BOVW Textual: GloVe, Vadar	Text model: aggregation score between (CNN and Senticircle). Visual model: SVM then finally applying Boolean operations	Late fusion	Collected tweets using #CWC2019 on two social media sites Instagram and Twitter	Accuracy: 91%
Kumar and Garg [187]	Visual: R-CNN, Sentibank for extracting med-level feature Textual: BOW, Unigrams, PoS, Negation, Count of Emoticon, Count of elongated words, Count of capitalised words, length of message	Text model: aggregation score between (Gradient Boosting, SentiCircle) Visual model: hybrid of SentiBank and RCNN and finally the result obtaained by Aggregating the Sentiment Polarity and scores	Late fusion	Collected tweets for LGBT verdict of Indian Penal Court (IPC) section 377.	Accuracy: 91.32%
Liao et al. [188]	Visual: EfficientNet-b0 Textual: GloVe and text level graph network (GNN) to extract contextualised features	Soft-max Classifier	Joint fusion	MVSA-single Twitter26k	Max accuracy: 0.7384 Max accuracy: 0.9388
Yu et al. [189]	Visual: CNN Textual: Word2vec	Logistic regression for textual and visual separately then apply average strategy for late fusion	Late fusion	Sina Weibo dataset	Accuracy for 2 class: 0.811 Accuracy for 3 class: 0.748

Y. Zhang et al. [190]	Visual: SIFT descriptor used for extracting the visual features to construct the BOVW Textual: GloVe	(RF), SVM then applying Quantum-inspired Multimodel Sentiment Analysis (QMSA) for fusion	Late fusion	Getty Images Flicker	Accuracy for RF: 0.8824 Accuracy for SVM: 0.7976 Accuracy for RF: 0.9314 Accuracy for SVM: 0.9243
Liu et al. [191]	Visual: VGG16 for extracting the sequence level feature concatenated with C3D to extract the frame level features Textual: Synset Forest procedure	Visual model: CONVLSTM then Softmax Layer Textual model: SentiWord-Net3.0 model then apply weighted sum for fusion	Late fusion	T-GIF Dataset: GSO-2016 Dataset Adjusted-GIFGIF Dataset	Accuracy: 0.7839 Accuracy: 0.7513 Accuracy: 0.7403
Qian et al. [192]	Visual: AlexNet Textual: AffectiveS-pace	Visual model: Softmax Layer Textual model: SVM	Late fusion	Twitter dataset	Accuracy: 0.8051
S. Zhang et al. [160]	Visual: VGG16 Textual: GloVe, FastText and Bert	Softmax classifier for joint fusion then a NN for late fusion	Combining Joint fusion and Late fusion	MM Yelp dataset CMU-MOSI and CMU-MOSEI Twitter-15 and Twitter-17	Avg. Accuracy: 63.41% Accuracy: 35.42%, 51.65% Accuracy: 78.62%, 71.35%
Quoc-Tuan Truong and Hady W. Lauw. [193]	Visual: VGG16 Textual: GloVe.	softmax classifier	Joint fusion	MM Yelp dataset	Avg. Accuracy: 61.88%
P. Kumar et al. [194]	Visual: VGG16. Textual: GloVe	Softmax classifier for joint fusion then a weighed average for late fusion	Combining joint and late fusion	Emotion Dataset B-T4SA dataset	Accuracy: 90.20% Accuracy: 86.70%

Declarations

- *Funding:* No funds, grants, or other support was received.
- *Conflict of interest:* The authors have no conflicts of interest to declare that are relevant to the content of this article.
- *Availability of data and materials:* Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

[1] Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69(1):14–23, 2014.

[2] Vadim Kagan, Andrew Stevens, and V.S. Subrahmanian. Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election. *IEEE Intelligent Systems*, 30(1):2–5, jan 2015.

[3] Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System. Technical report.

[4] Yannis Panagakis, Jean Kossaifi, Grigorios G. Chrysos, James Oldfield, Mihalis A. Nicolaou, Anima Anandkumar, and Stefanos Zafeiriou. Tensor Methods in Computer Vision and Deep Learning. *Proceedings of the IEEE*, 109(5):863–890, may 2021.

[5] Jianbo Yu, Chengyi Zhang, and Shijin Wang. Multichannel one-dimensional convolutional neural network-based feature learning for fault diagnosis of industrial processes. *Neural Computing and Applications*, 33(8):3085–3104, apr 2021.

[6] Dejun Zhang, Long Tian, Mingbo Hong, Fei Han, Yafeng Ren, and Yilin Chen. Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification. *IEEE Access*, 6:73750–73759, 2018.

[7] Mingyuan Xin and Yong Wang. Research on image classification model based on deep convolution neural network. *Eurasip Journal on Image and Video Processing*, 2019(1), dec 2019.

[8] Hongli Zhang, Alireza Jolfaei, and Mamoun Alazab. A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing. *IEEE Access*, 7:159081–159089, 2019.

[9] Ahammed M S Shaik Afzal. Optimized Support Vector Machine Model for Visual Sentiment Analysis. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pages 171–175, may 2021.

[10] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. Can we understand van gogh's mood? In *Proceedings of the 20th ACM international conference on Multimedia - MM '12*, page 857, New York, New York, USA, 2012. ACM Press.

[11] Mayank Amencherla and Lav R. Varshney. Color-based visual sentiment for social communication. In *2017 15th Canadian Workshop on Information Theory (CWIT)*, pages 1–5, jun 2017.

[12] Lifang Wu, Mingchao Qi, Meng Jian, and Heng Zhang. Visual Sentiment Analysis by Combining Global and Local Information. *Neural Processing Letters*, 51(3):2063–2075, jun 2020.

[13] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. How Do Your Friends on Social Media Disclose Your Emotions? In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 306–312, 2014.

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

[15] Brendan Jou, Subhabrata Bhattacharya, and Shih Fu Chang. Predicting viewer perceived emotions in animated GIFs. In *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pages 213–216. Association for Computing Machinery, nov 2014.

[16] N. Desai, S. Venkatramana, and B. V.D.S. Sekhar. Automatic visual sentiment analysis with convolution neural network. *International Journal of Industrial Engineering and Production Research*, 31(3):351–360, 2020.

[17] Jie Chen, Qirong Mao, and Luoyang Xue. Visual sentiment analysis with active learning. *IEEE Access*, 8:185899–185908, 2020.

[18] Kaikai Song, Ting Yao, Qiang Ling, and Tao Mei. Boosting image sentiment analysis with visual attention. *Neurocomputing*, 312:218–228, oct 2018.

[19] Zhuanghui Wu, Min Meng, and Jigang Wu. Visual Sentiment Prediction with Attribute Augmentation and Multi-attention Mechanism. *Neural Processing Letters*, 51(3):2403–2416, jun 2020.

[20] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. A Deep Learning Perspective on Beauty, Sentiment, and Remembrance of Art. *IEEE Access*, 7:73694–73710, 2019.

[21] Jufeng Yang, Dongyu She, Ming Sun, Ming Ming Cheng, Paul L. Rosin, and Liang Wang. Visual Sentiment Prediction Based on Automatic Discovery of Affective Regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, sep 2018.

[22] Haochun Ou, Chunmei Qing, Xiangmin Xu, and Jianxiu Jin. Multi-level context pyramid network for visual sentiment analysis. *Sensors*, 21(6):1–20, mar 2021.

[23] Ashima Yadav and Dinesh Kumar Vishwakarma. A deep learning architecture of RA-DLNet for visual sentiment analysis. *Multimedia Systems*, 26(4):431–451, aug 2020.

[24] Ziyuan Zhao, Huiying Zhu, Z. Xue, Zhao Liu, Jing Tian, Matthew Chin Heng Chua, and M. Liu. An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing and Management*, 56(6), nov 2019.

[25] Xia Li and Minping Chen. Multimodal Sentiment Analysis with Multi-perspective Fusion Network Focusing on Sense Attentive Language. In *Chinese Computational Linguistics*, pages 359–373. 2020.

[26] Agnieszka Rozanska and Michal Podpora. Multimodal sentiment analysis applied to interaction between patients and a humanoid robot Pepper. In *IFAC-PapersOnLine*, volume 52, pages 411–414, 2019.

[27] Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, dec 2005.

[28] Julien Deonna and Fabrice Teroni. *The Emotions*. Routledge, nov 2012.

[29] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111, 2014.

[30] Chetashri Bhadane, Hardi Dalal, and Heenal Doshi. Sentiment analysis: Measuring opinions. In *Procedia Computer Science*, volume 45, pages 808–814. Elsevier B.V., 2015.

[31] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers, 2018.

[32] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of Twitter sentiment analysis methods, jun 2016.

[33] Bing Liu. *Sentiment analysis and opinion mining*. Morgan & Claypool, 2012.

[34] Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence*, 48(5):1218–1232, may 2018.

[35] Amit Kumar Sharma, Sandeep Chaurasia, and Devesh Kumar Srivastava. Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec. In *Procedia Computer Science*, volume 167, pages 1139–1147. Elsevier B.V., 2020.

[36] Kashfia Sailunaz and Reda Alhajj. Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, 36, sep 2019.

[37] Saurav Pradha, Malka N. Halgamuge, and Nguyen Tran Quoc Vinh. Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–8. IEEE, oct 2019.

[38] Mohammed Kasri, Marouane Birjali, and Abderrahim Beni-Hssane. A comparison of features extraction methods for Arabic sentiment analysis. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, oct 2019.

[39] M. Avinash and E. Sivasankar. A study of feature extraction techniques for sentiment analysis. In *Advances in Intelligent Systems and Computing*, volume 814, pages 475–486. Springer Verlag, 2019.

[40] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, aug 2013.

[41] Kevin Gimpel, Nathan Schneider, Brendan O'connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[42] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, volume 1, pages 173–180, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[43] Vishal A. and S.S. Sonawane. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications*, 139(11):5–15, apr 2016.

[44] Charu C Aggarwal and Chengxiang Zhai. *Mining Text Data*. Springer US, Boston, MA, 2012.

[45] Vincenzo Loia and Sabrina Senatore. A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content. *Knowledge-Based Systems*, 58:75–85, mar 2014.

[46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, jan 2013.

[47] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

[48] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, aug 2020.

[49] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, dec 2017.

[50] Mohammed Jabreel and Antonio Moreno. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9(6), 2019.

[51] Ibrahim Kaibi, El Habib Nfaoui, and Hassan Satori. A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis. In *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, pages 1–4. IEEE, apr 2019.

[52] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. The impact of features extraction on the sentiment analysis. In *Procedia Computer Science*, volume 152, pages 341–348. Elsevier B.V., 2019.

[53] Doaa Mohey. Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, 7(1), 2016.

[54] Alexander Mee, Elmina Homapour, Francisco Chiclana, and Ofer Engel. Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit[Formula presented]. *Knowledge-Based Systems*, 228, sep 2021.

[55] M. D. Devika, C. Sunitha, and Amal Ganesh. Sentiment Analysis: A Comparative Study on Different Approaches. In *Procedia Computer Science*, volume 87, pages 44–49. Elsevier B.V., 2016.

[56] Nor Nadiah Yusof, Azlinah Mohamed, and Shuzlina Abdul-Rahman. Reviewing classification approaches in sentiment analysis. In *Communications in Computer and Information Science*, volume 545, pages 43–53. Springer Verlag, 2015.

[57] Luca Oneto, Federica Bisio, Erik Cambria, and Davide Anguita. Statistical Learning Theory and ELM for Big Social Data Analysis. *IEEE Computational Intelligence Magazine*, 11(3):45–55, aug 2016.

[58] Yang Li, Quan Pan, Tao Yang, Suhang Wang, Jiliang Tang, and Erik Cambria. Learning Word Representations for Sentiment Analysis. *Cognitive Computation*, 9(6):843–851, dec 2017.

[59] Basant Agarwal and Namita Mittal. *Prominent Feature Extraction for Sentiment Analysis*. Socio-Affective Computing. Springer International Publishing, Cham, 2016.

[60] Gwanghoon Yoo and Jeesun Nam. A Hybrid Approach to Sentiment Analysis Enhanced by Sentiment Lexicons and Polarity Shifting Devices - Archive ouverte HAL. In *the 13th Workshop on asian Language Resources*, pages 21–28, 2018.

[61] Dominik Fisch, Edgar Kalkowski, and Bernhard Sick. Knowledge fusion for probabilistic generative classifiers with data mining applications. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):652–666, mar 2014.

[62] Zhi Li, Rui Li, and Guanghao Jin. Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary. *IEEE Access*, 8:75073–75084, 2020.

[63] Ankit Shrivastava, Onkar Kshire, Komal Pipriya, Rishika Shivhare, and Uma Thakur. SENTIMENT ANALYSIS USING NAÏVE BAYE'S ALGORITHM. *International Research Journal of Modernization in Engineering Technology and Science*, 3398:2582–5208, 2021.

[64] Sourav Kunal, Arijit Saha, Aman Varma, and Vivek Tiwari. Textual Dissection of Live Twitter Reviews using Naive Bayes. In *Procedia Computer Science*, volume 132, pages 307–313. Elsevier B.V., 2018.

[65] Rizqi Sukma Kharisma, Muttafi'ah, and Akhmad Dahlan. Comparison of Naïve Bayes Algorithm Model Combinations with Term Weighting Techniques in Sentiment Analysis. *ICOIACT 2021 - 4th International Conference on Information and Communications Technology: The Role of AI in Health and Social Revolution in Turbulence Era*, pages 160–163, 2021.

[66] Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. Supervised sentiment analysis in Czech social media. *Information Processing and Management*, 50(5):693–707, 2014.

[67] Hein Htet and Yi Yi Myint. Social media (Twitter) Data analysis using maximum entropy classifier on big data processing framework (Case study: Analysis of health condition, education status, states of business). *~695 ~ Journal of Pharmacognosy and Phytochemistry*, 1:695–700, 2018.

[68] Pierre Ficamos, Yan Liu, and Weiyi Chen. A Naive Bayes and Maximum Entropy approach to sentiment analysis: Capturing domain-specific data in Weibo. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 336–339. IEEE, feb 2017.

[69] Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. Webis: An Ensemble for Twitter Sentiment Detection. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 582–589, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

[70] Luis Gutiérrez, Juan Bekios-Calfa, and Brian Keith. A review on bayesian networks for sentiment analysis. In *Advances in Intelligent Systems and Computing*, volume 865, pages 111–120. Springer Verlag, 2019.

[71] Yun Wan and Qigang Gao. An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. In *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, pages 1318–1325. Institute of Electrical and Electronics Engineers Inc., jan 2016.

[72] Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. *Journal of Computational Science*, 27:386–393, jul 2018.

[73] Nur Syafiqah Mohd Nafis and Suryanti Awang. An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification. *IEEE Access*, 9:52177–52192, 2021.

[74] Ruba Obiedat, Raneem Qaddoura, Ala'M M. Al-Zoubi, Laila Al-Qaisi, Osama Harfoushi, Mo'ath Alrefai, and Hossam Faris. Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution. *IEEE Access*, 10:22260–22273, 2022.

[75] Mohammad Rezwanul, Ahmad Ali, and Anika Rahman. Sentiment Analysis on Twitter Data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, 8(6), 2017.

[76] Sheeba Naz, Aditi Sharan, and Nidhi Malik. Sentiment Classification on Twitter Data Using Support Vector Machine. In *Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018*, pages 676–679. Institute of Electrical and Electronics Engineers Inc., jan 2019.

[77] Naw Naw. Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers. *International Journal of Scientific and Research Publications (IJSRP)*, 8(10), oct 2018.

[78] M. Rizzo Irfan, M. Ali Fauzi, Tibyani Tibyani, and Nurul Dyah Mentari. Twitter Sentiment Analysis on 2013 Curriculum Using Ensemble Features and K-Nearest Neighbor. *International Journal of Electrical and Computer Engineering (IJECE)*, 8(6):5409, dec 2018.

[79] R. Damarta, A. Hidayat, and A. S. Abdullah. The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality. In *Journal of Physics: Conference Series*, volume 1722. IOP Publishing Ltd, jan 2021.

[80] Elik Hari Muktafin and Pramono Kusrini. Sentiments analysis of customer satisfaction in public services using K-nearest neighbors algorithm and natural language processing approach. *Telkomnika (Telecommunication Computing Electronics and Control)*, 19(1):146–154, feb 2021.

[81] Vo Ngoc Phu, Vo Thi Ngoc Tran, Vo Thi Ngoc Chau, Nguyen Duy Dat, and Khanh Ly Doan Duy. A decision tree using ID3 algorithm for English semantic analysis. *International Journal of Speech Technology*, 20(3):593–613, sep 2017.

[82] Fatima Es-Sabery, Khadija Es-Sabery, Junaid Qadir, Beatriz Sainz-De-Abajo, Abdellatif Hair, Begoña García-Zapirain, and Isabel De La Torre-Díez. A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier. *IEEE Access*, 9:58706–58739, 2021.

[83] Veny Amilia Fitri, Rachmadita Andreswari, and Muhammad Azani Hasibuan. Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. In *Procedia Computer Science*, volume 161, pages 765–772. Elsevier B.V., 2019.

[84] Rupal Bhargava, Shivangi Arora, and Yashvardhan Sharma. Neural network-based architecture for sentiment analysis in Indian languages. *Journal of Intelligent Systems*, 28(3):361–375, jul 2019.

[85] Rodrigo Moraes, João Francisco Valiati, and Wilson P. Gavião Neto. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2):621–633, feb 2013.

[86] G. Vinodhini and R. M. Chandrasekaran. A comparative performance evaluation of neural network based approach for sentiment classification of online reviews. *Journal of King Saud University - Computer and Information Sciences*, 28(1):2–12, jan 2016.

[87] B. S. Harish, Bhanu Prasad, and B. Udayasri. Classification of text documents using adaptive fuzzy C-means clustering. *Advances in Intelligent Systems and Computing*, 235:205–214, 2014.

[88] Korawit Orkphol and Wu Yang. Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony. *International Journal of Computational Intelligence and Applications*, 18(3), sep 2019.

[89] Yue Han, Yuhong Liu, and Zhigang Jin. Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. *Neural Computing and Applications*, 32(9):5117–5129, may 2020.

[90] Milagros Fernández-Gavilanes, Tamara Álvarez-López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58:57–75, oct 2016.

[91] Aitor García-Pablos, Montse Cuadros, and German Rigau. W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, 91:127–137, jan 2018.

[92] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database [*]. *International Journal of Lexicography*, 3(4):235–244, 1990.

[93] Saif Mohammad, Cody Dunne, and Bonnie Dorr. 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 Volume 2 - EMNLP '09*, volume 2, page 599, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[94] Minghui Huang, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard K.M. Poon, and Fu Lee Wang. Lexicon-Based Sentiment Convolutional Neural Networks for Online Review Analysis. *IEEE Transactions on Affective Computing*, 2020.

[95] Liang Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(3):671–681, mar 2018.

[96] Anna Jurek, Maurice D. Mulvenna, and Yaxin Bi. Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(1), dec 2015.

[97] Swati Sanagar and Deepa Gupta. Unsupervised Genre-Based Multidomain Sentiment Lexicon Learning Using Corpus-Generated Polarity Seed Words. *IEEE Access*, 8:118050–118071, 2020.

[98] Seongik Park and Yanggon Kim. Building thesaurus lexicon using dictionary-based approach for sentiment classification. In *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 39–44. IEEE, jun 2016.

[99] Swati Sanagar and Deepa Gupta. Roadmap for polarity lexicon learning and resources: A survey. In *Advances in Intelligent Systems and Computing*, volume 530, pages 647–663. Springer Verlag, 2016.

[100] Gang Wang, Zhu Zhang, Jianshan Sun, Shanlin Yang, and Catherine A. Larson. POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis. *Information Processing and Management*, 51(4):458–479, jul 2015.

[101] Vandana Jha, R. Savitha, P. Deepa Shenoy, Venugopal K R, and Arun Kumar Sangaiah. A novel sentiment aware dictionary for multi-domain sentiment classification. *Computers and Electrical Engineering*, 69:585–597, jul 2018.

[102] Basant Agarwal and Namita Mittal. Semantic Orientation-Based Approach for Sentiment Analysis. In *Prominent Feature Extraction for Sentiment Analysis*, pages 77–88. 2016.

[103] Tiejian Luo, Su Chen, Guandong Xu, and Jia Zhou. Sentiment Analysis. In *Trust-based Collective View Prediction*, pages 53–68. Springer New York, New York, NY, 2013.

[104] Abdallah Yousif, Zhendong Niu, John K. Tarus, and Arshad Ahmad. A survey on sentiment analysis of scientific citations. *Artificial Intelligence Review*, 52(3):1805–1838, oct 2019.

[105] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications*, 125(3):26–33, sep 2015.

[106] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266–6282, 2013.

[107] Farhan Hassan Khan, Saba Bashir, and Usman Qamar. TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57(1):245–257, jan 2014.

[108] D. V.Nagarjuna Devi, Thatiparti Venkata Rajini Kanth, Kakollu Mounika, and Nambhatla Sowjanya Swathi. Assay: Hybrid approach for sentiment analysis. In *Smart Innovation, Systems and Technologies*, volume 106, pages 309–318. Springer Science and Business Media Deutschland GmbH, 2019.

[109] Bonggun Shin, Timothy Lee, and Jinho D. Choi. Lexicon Integrated CNN Models with Attention for Sentiment Analysis, oct 2016.

[110] Kariman Elshakankery and Mona F. Ahmed. HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis. *Egyptian Informatics Journal*, 20(3):163–171, nov 2019.

[111] Lina Maria Rojas-Barahona. Deep learning for sentiment analysis. *Language and Linguistics Compass*, 10(12):701–719, dec 2016.

[112] Peerapon Vateekul and Thanabhat Koomsubha. 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)*, pages 1–6. IEEE, jul 2016.

[113] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), jul 2018.

[114] Abdulaziz M. Alayba and Vasile Palade. Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation. *Journal of King Saud University - Computer and Information Sciences*, 2021.

[115] Mehmet Umut Salur and Ilhan Aydin. A Novel Hybrid Deep Learning Model for Sentiment Classification. *IEEE Access*, 8:58080–58093, 2020.

[116] Muhammad Zulqarnain, Suhaimi Abd Ishak, Rozaida Ghazali, Nazri Mohd Nawi, Muhammad Aamir, and Yana Mazwin Mohmad Hassim. An improved deep learning approach based on variant two-state gated recurrent unit and word embeddings for sentiment classification. *International Journal of Advanced Computer Science and Applications*, 11(1):594–603, 2020.

[117] Liangqiang Li, Liang Yang, and Yuyang Zeng. Improving sentiment classification of restaurant reviews with attention-based bi-gru neural network. *Symmetry*, 13(8), aug 2021.

[118] Amit Kumar Sharma, Sandeep Chaurasia, and Devesh Kumar Srivastava. Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec. In *Procedia Computer Science*, volume 167, pages 1139–1147. Elsevier B.V., 2020.

[119] Mohammad Ehsan Basiri, Moloud Abdar, Mehmet Akif Cifci, Shahla Nemati, and U. Rajendra Acharya. A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques. *Knowledge-Based Systems*, 198, jun 2020.

[120] Wei Li, Luyao Zhu, Yong Shi, Kun Guo, and Erik Cambria. User reviews: Sentiment analysis using lexicon integrated two-channel CNN–LSTM family models. *Applied Soft Computing Journal*, 94, sep 2020.

[121] Fazeel Abid, Muhammad Alam, Muhammad Yasir, and Chen Li. Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. *Future Generation Computer Systems*, 95:292–308, jun 2019.

[122] Cach N. Dang, María N. Moreno-García, and Fernando De La Prieta. Hybrid Deep Learning Models for Sentiment Analysis. *Complexity*, 2021, 2021.

[123] Masoud Fatemi and Mehran Safayani. Joint sentiment/topic modeling on text data using a boosted restricted Boltzmann Machine. *Multimedia Tools and Applications*, 78(15):20637–20653, aug 2019.

[124] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 5733–5742. IEEE Computer Society, dec 2016.

[125] Alramzana Nujum Navaz, Serhani Mohamed Adel, and Sujith Samuel Mathew. Facial Image Pre-Processing and Emotion Classification: A Deep Learning Approach. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE, nov 2019.

[126] D. Tamil Priya and J. Divya Udayan. Transfer learning techniques for emotion classification on visual features of images in the deep learning network. *International Journal of Speech Technology*, 23(2):361–372, jun 2020.

[127] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia - MM '10*, page 83, New York, New York, USA, 2010. ACM Press.

[128] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, nov 1991.

[129] Markus A. Stricker and Markus Orengo. Similarity of color images. In *Storage and Retrieval for Image and Video Databases III*, volume 2420, page 381. SPIE, mar 1995.

[130] G. Pass and R. Zabih. Histogram refinement for content-based image retrieval. In *Proceedings Third IEEE Workshop on Applications of Computer Vision. WACV'96*, pages 96–102. IEEE Comput. Soc. Press, 1996.

[131] John R. Smith and Shih-Fu Chang. Tools and techniques for color image retrieval. In *Storage and Retrieval for Still Image and Video Databases IV*, volume 2670, pages 426–437. SPIE, mar 1996.

[132] Jing Huang. *COLOR-SPATIAL IMAGE INDEXING AND APPLICATIONS*. PhD thesis, Cornell University, 1998.

[133] Anupam Mukherjee, Kumar Gaurav, Abhishek Verma, Harishankar Kumar, and Rahul Thakur. Content Based Image Retrieval using GLCM Predictive Crime Data Analysis View project Content Based Image Retrieval using GLCM. *Article in International Journal of Innovative Research in Computer and Communication Engineering*, 3297(11), 2016.

[134] Savita, Sandeep Jain, and K.K. Paliwal. CBIR by cascading features & SVM. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 93–97. IEEE, may 2017.

[135] Pavel Vacha, Michal Haindl, and Tomas Suk. Colour and rotation invariant textural features based on Markov random fields. *Pattern Recognition Letters*, 32(6):771–779, apr 2011.

[136] T.N. Tan and A.G. Constantinides. Texture feature extraction based on primitive analysis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1683–1686. IEEE, 1989.

[137] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, jul 1989.

[138] Xin Shu, Lei Pan, and Xiao Jun Wu. Multi-scale contour flexibility shape signature for Fourier descriptor. *Journal of Visual Communication and Image Representation*, 26:161–167, 2015.

[139] Paul V C Hough. METHOD AND MEANS FOR RECOGNIZING COMPLEX PATTERNS, 1962.

[140] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IEEE Transactions on Information Theory*, 8(2):179–187, feb 1962.

[141] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. In *Computer Vision - ECCV 2000*, pages 404–420, 2000.

[142] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, pages 223–232. Association for Computing Machinery, 2013.

[143] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*, pages 1–8, New York, New York, USA, 2013. ACM Press.

[144] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan Ying Chen, and Shih Fu Chang. Object-based visual sentiment concept analysis and application. In *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pages 367–376. Association for Computing Machinery, nov 2014.

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

[146] Stephen Pilli, Manasi Patwardhan, Niranjan Pedanekar, and Shirish Karande. Predicting sentiments in image advertisements using semantic relations among sentiment labels. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2020-June, pages 1640–1648. IEEE Computer Society, jun 2020.

[147] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, sep 2014.

[148] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807. IEEE, jul 2017.

[149] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, jun 2016.

[150] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, jun 2016.

[151] Henrik Boström, Sten F Andler, Marcus Brohede, Ronnie Johansson, Alexander Karlsson, Joeri Van Laere, Lars Niklasson, Maria Nilsson, Anne Persson, and Tom Ziemke. On the Definition of Information Fusion as a Field of Research. Technical report, 2007.

[152] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, mar 1998.

[153] Akshi Kumar, Kathiravan Srinivasan, Wen Huang Cheng, and Albert Y. Zomaya. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing and Management*, 57(1), jan 2020.

[154] Kang Zhang, Yushui Geng, Jing Zhao, Jianxin Liu, and Wenxiao Li. Sentiment Analysis of Social Media via Multimodal Feature Fusion. *Symmetry*, 12(12):2010, dec 2020.

[155] Ke Zhang, Yunwen Zhu, Wenjun Zhang, and Yonghua Zhu. Cross-modal image sentiment analysis via deep correlation of textual semantic. *Knowledge-Based Systems*, 216, mar 2021.

[156] Meng Cao, Yonghua Zhu, Wenjing Gao, Mengyao Li, and Shaoxiu Wang. Various syncretic co-attention network for multimodal sentiment analysis. *Concurrency and Computation: Practice and Experience*, 32(24), dec 2020.

[157] Jie Xu, Zhoujun Li, Feiran Huang, Chaozhuo Li, and Philip S. Yu. Social Image Sentiment Analysis by Exploiting Multimodal Content and Heterogeneous Relations. *IEEE Transactions on Industrial Informatics*, 17(4):2974–2982, apr 2021.

[158] Jie Xu, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowledge-Based Systems*, 178:61–73, 2019.

[159] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

[160] Sun Zhang, Bo Li, and Chunyong Yin. Cross-modal sentiment sensing with visual-augmented representation and diverse decision fusion. *Sensors*, 22(1), jan 2022.

[161] Feiran Huang, Kaimin Wei, Jian Weng, and Zhoujun Li. Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 16(3), sep 2020.

[162] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated Multimodal Units for Information Fusion, feb 2017.

[163] Joshua B. Tenenbaum and William T. Freeman. Separating Style and Content with Bilinear Models. *Neural Computation*, 12(6):1247–1283, jun 2000.

[164] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Stroudsburg, PA, USA, jul 2017. Association for Computational Linguistics.

[165] Jin Ye, Xiaojiang Peng, Yu Qiao, Hao Xing, Junli Li, and Rongrong Ji. Visual-Textual Sentiment Analysis in Product Reviews. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 869–873. IEEE, sep 2019.

[166] Tao Zhou, Jiuxin Cao, Xuelin Zhu, Bo Liu, and Shancang Li. Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction. *IEEE Systems Journal*, 15(3):4303–4314, oct 2020.

[167] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1008–1017, New York, NY, USA, oct 2016. ACM.

[168] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. Multimodal Sentiment Analysis With Image-Text Interaction Network. *IEEE Transactions on Multimedia*, pages 1–1, 2022.

[169] Nan Xu, Wenji Mao, and Guandan Chen. A co-memory network for multimodal sentiment analysis. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pages 929–932. Association for Computing Machinery, Inc, jun 2018.

[170] Tao Jiang, Jiahai Wang, Zhiyue Liu, and Yingbiao Ling. Fusion-Extraction Network for Multimodal Sentiment Analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12085 LNAI, pages 785–797. Springer, 2020.

[171] Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan. LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition. *IEEE Transactions on Multimedia*, 23:1785–1798, 2021.

[172] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, pages 13–22. Association for Computing Machinery, Inc, feb 2016.

[173] Chi Thang Duong, Remi Lebret, and Karl Aberer. Multimodal Classification for Analysing Social Media, aug 2017.

[174] Claudio Baecchi, Tiberio Uricchio, Marco Bertini, and Alberto Del Bimbo. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications*, 75(5):2507–2525, mar 2016.

[175] Cheng Peng, Chunxia Zhang, Xiaojun Xue, Jiameng Gao, Hongjian Liang, and Zhengdong Niu. Cross-modal complementary network with hierarchical fusion for multimodal sentiment classification. *Tsinghua Science and Technology*, 27(4):664–679, aug 2022.

[176] Rongrong Ji, Fuhai Chen, Liujuan Cao, and Yue Gao. Cross-Modality Microblog Sentiment Prediction via Bi-Layer Multimodal Hypergraph Learning. *IEEE Transactions on Multimedia*, 21(4):1062–1075, apr 2019.

[177] Jie Xu, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. Sentiment analysis of social images via hierarchical deep fusion of content and links. *Applied Soft Computing Journal*, 80:387–399, jul 2019.

[178] Tsegaye Misikir Tashu, Sakina Hajiyeva, and Tomas Horvath. Multimodal emotion recognition from art using sequential co-attention. *Journal of Imaging*, 7(8), aug 2021.

[179] Alessandro Ortis, Giovanni Maria Farinella, Giovanni Torrisi, and Sebastiano Battiato. Exploiting objective text description of images for visual sentiment analysis. *Multimedia Tools and Applications*, 80(15):22323–22346, jun 2021.

[180] Guoyong Cai and Binbin Xia. Convolutional Neural Networks for Multimedia Sentiment Analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9362, pages 159–167. Springer Verlag, 2015.

[181] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026, 2021.

[182] Ashima Yadav and Dinesh Kumar Vishwakarma. A Deep Multi-Level Attentive network for Multimodal Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, mar 2022.

[183] Hang Miao, Ruifang Liu, Sheng Gao, Xin Zhou, and Xiaoxin He. End-to-End Deep Memory Network for Visual-Textual Sentiment Analysis. In *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 399–403. IEEE, aug 2018.

[184] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167:26–37, mar 2019.

[185] Xuelin Zhu, Biwei Cao, Shuai Xu, Bo Liu, and Jiuxin Cao. Joint Visual-Textual Sentiment Analysis Based on Cross-Modality Attention Mechanism. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11295 LNCS, pages 264–276. Springer Verlag, 2019.

[186] Nan Xu and Wenji Mao. MultiSentiNet: A deep semantic network for multimodal sentiment analysis. In *International Conference on Information and Knowledge Management, Proceedings*, volume Part F131841, pages 2399–2402. Association for Computing Machinery, nov 2017.

[187] Akshi Kumar and Geetanjali Garg. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications*, 2019.

[188] Wenxiong Liao, Bi Zeng, Jianqi Liu, Pengfei Wei, and Jiongkun Fang. Image-text interaction graph neural network for image-text sentiment analysis. *Applied Intelligence*, jan 2022.

[189] Yuhai Yu, Hongfei Lin, Jiana Meng, and Zhehuan Zhao. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms*, 9(2), jun 2016.

[190] Yazhou Zhang, Dawei Song, Peng Zhang, Panpan Wang, Jingfei Li, Xiang Li, and Benyou Wang. A quantum-inspired multimodal sentiment analysis framework. *Theoretical Computer Science*, 752:21–40, dec 2018.

[191] Tianliang Liu, Junwei Wan, Xiubin Dai, Feng Liu, Quanzeng You, and Jiebo Luo. Sentiment Recognition for Short Annotated GIFs Using Visual-Textual Fusion. *IEEE Transactions on Multimedia*, 22(4):1098–1110, apr 2020.

[192] Chen Qian, Edoardo Ragusa, Iti Chaturvedi, Erik Cambria, and Rodolfo Zunino. Text-Image Sentiment Analysis. Technical report.

[193] Quoc-Tuan Truong and Hady W. Lauw. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:305–312, jul 2019.

[194] Puneet Kumar, Vedanti Khokher, Yukti Gupta, and Balasubramanian Raman. HYBRID FUSION BASED APPROACH FOR MULTIMODAL EMOTION RECOGNITION WITH INSUFFICIENT LABELED DATA. In *Proceedings - International Conference on Image Processing, ICIP*, volume 2021-Septe, pages 314–318. IEEE Computer Society, 2021.

[195] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment Analysis on Multi-View Social Data. In Qi Tian, Nicu Sebe, Guo-Jun Qi, Benoit Huet, Richang Hong, and Xueliang Liu, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9517 of *Lecture Notes in Computer Science*, pages 15–27. Springer International Publishing, Cham, 2016.

[196] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-Media Learning for Image Sentiment Analysis in the Wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317. IEEE, oct 2017.

[197] Mauro Dragoni and Giulio Petrucci. A fuzzy-based strategy for multi-domain sentiment analysis. *International Journal of Approximate Reasoning*, 93:59–73, feb 2018.

[198] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N H Shwetha. An improved sentiment analysis of online movie reviews based on clustering for box-office prediction. In *International Conference on Computing, Communication & Automation*, pages 933–937. IEEE, may 2015.

[199] Heena Timani, Parag Shah, and Mansi Joshi. Predicting Success of a Movie from Youtube Trailer Comments using Sentiment Analysis | IEEE Conference Publication | IEEE Xplore. In *th International Conference on Computing for Sustainable Global Development*, 2019.

[200] Amin Hedayati Moghaddam, Moein Hedayati Moghaddam, and Morteza Esfandyari. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, 21(41):89–93, dec 2016.

[201] Prakash P. Rokade and D. Aruna Kumari. Business intelligence analytics using sentiment analysis-a survey. *International Journal of Electrical and Computer Engineering*, 9(1):613–620, feb 2019.

[202] Joseph G. Ellis, Brendan Jou, and Shih Fu Chang. Why we watch the news: A dataset for exploring sentiment in broadcast video news. In *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, pages 104–111. Association for Computing Machinery, Inc, nov 2014.

[203] Xiao Lin Zheng, Chao Chao Chen, Jui Long Hung, Wu He, Fu Xing Hong, and Zhen Lin. A Hybrid Trust-Based Recommender System for Online Communities of Practice. *IEEE Transactions on Learning Technologies*, 8(4):345–356, oct 2015.

[204] Priyavrat Chauhan, Nonita Sharma, and Geeta Sikka. The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 12(2):2601–2627, feb 2021.

[205] Francisco Javier Ramírez-Tinoco, Giner Alor-Hernández, José Luis Sánchez-Cervantes, María del Pilar Salas-Zárate, and Rafael Valencia-García. Use of Sentiment Analysis Techniques in Healthcare Domain. *Studies in Computational Intelligence*, 815:189–212, 2019.