Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions

Article *in* Behaviour and Information Technology · December 2022

DOI: 10.1080/0144929X.2022.2156387

2 authors:

Alaa Alslaity
Dalhousie University
29 PUBLICATIONS **98** CITATIONS
SEE PROFILE

Rita Orji
Dalhousie University
287 PUBLICATIONS **5,628** CITATIONS
SEE PROFILE

Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions

Alaa Alslaity & Rita Orji

To cite this article: Alaa Alslaity & Rita Orji (2022): Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions, Behaviour & Information Technology, DOI: 10.1080/0144929X.2022.2156387

To link to this article: https://doi.org/10.1080/0144929X.2022.2156387

Published online: 16 Dec 2022.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2022.2156387

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions

Alaa Alslaity and Rita Orji

Faculty of Computer Science, Dalhousie University, Halifax, Canada

ABSTRACT

Emotion detection and Sentiment analysis techniques are used to understand polarity or emotions expressed by people in many cases, especially during interactive systems use. Recognizing users' emotions is an important topic for human–computer interaction. Computers that recognize emotions would provide more natural interactions. Also, emotion detection helps design human-centred systems that provide adaptable behaviour change interventions based on users' emotions. The growing capability of machine learning to analyze big data and extract emotions therein has led to a surge in research in this domain. With this increased attention, it becomes essential to investigate this research area and provide a comprehensive review of the current state. In this paper, we conduct a systematic review of 123 papers on machine learning-based emotion detection to investigate research trends along many themes, including machine learning approaches, application domain, data, evaluation, and outcome. The results demonstrate: 1) increasing interest in this domain, 2) supervised machine learning (namely, SVM and Naïve Bayes) are the most popular algorithms, 3) Text datasets in the English language are the most common data source, and 4) most research use Accuracy to evaluate performance. Based on the findings, we suggest future directions and recommendations for developing human-centred systems.

ARTICLE HISTORY
Received 26 March 2022
Accepted 2 December 2022

ADDITIONAL KEYWORDS AND PHRASES
Emotion detection; emotion recognition; sentiment analysis; opinion mining; machine learning; deep learning

1. Introduction

In human-human interactions, it is necessary to know others' emotions and feelings in order to have better communication and add interactions accordingly. For instance, people tend to avoid negative conversations with somebody who feels anger. On the other hand, a salesperson may adjust the price if customers are excited to buy the product. These human-human interaction principles can be applied to computer–human interactions, although computers do not feel and have the same communication skills as humans (Yu, Benbasat, and Cenfetelli 2011). Furthermore, researchers have concluded that a computer system's interaction with humans is as important as its performance (Alslaity and Tran 2021). Therefore, understanding users' behaviour/reactions in response to interactive systems use is crucial to the success of any system, especially personalised and adaptive systems. This can be used to evaluate a system and understand usability and user experience issues using rich and diverse user reviews available online. The process of analyzing and detecting opinions and emotions expressed by people is called *Emotion Detection* (or *Sentiment Analysis*) (Saad 2014) and (Kaur and Saini 2014).

The domain of emotion detection has gained increasing attention for several reasons; firstly, the advances in approaches and techniques used to analyze data and detect emotions and opinions. Second, the growing popularity of technologies allowed and simplified the data collection process. These technologies have penetrated and integrated into our daily lives, making it possible to generate massive amounts of data about users and their behaviour from various sources. These resources include sensors, such as wearable devices and mobile phone equipment, and via the web, such as social media platforms, including Twitter and Facebook. This has led to the explosion of big data and the Internet of Behaviours (Javaid et al. 2021). Third, users' interest and engagement in these advanced technologies have grown significantly (Saad 2014), which, in turn, leads to an exponential increase in the available data.

This huge amount of data from various resources and in different forms is a rich source for understanding users' emotions and opinions, which can enhance systems' interactions with users. Emotion analysis can be used to inform adaptive systems design and could also be used to evaluate interactive systems and understand

CONTACT Alaa Alslaity alaa.alslaity@dal.ca

users' opinions and reaction design. For instance, users' reviewss on the Mobile eCommerce (Olagunju, Oyebode, and Orji 2020) apps have been used to explore issues hampering user experience. It has also been used to evaluate mental health apps to understand usability issues by applying machine learning on users' reviews available online (Oyebode, Alqahtani, and Orji 2020). However, extracting meaningful information and analyzing emotions from such unstructured data is not straightforward. One research direction that has made this possible is the advances in Machine Learning (ML) approaches.

Machine learning is a technique that concerns building models that learn automatically by experience (Naqa and Murphy 2015). The general idea of these approaches is to build a learning model using sample data, then make predictions, suggestions, or decisions without being given explicit instructions. ML approaches are broadly divided into three categories, Supervised, unsupervised, and semi-supervised learning. The key difference between these three categories is data labels; Supervised ML approaches learn from fully labelled data, while unsupervised ML learns from unlabelled data. Semi-supervised ML is an intermediate approach that uses partially labelled data to learn the unlabelled portion of the data. ML approaches have been applied successfully in diverse fields, such as pattern recognition, finance, learning, as well as emotion detection.

ML approaches show promising results in detecting users' emotions from several data sources, including text, speech, video, and more. Also, some ML approaches have shown high performance in processing and learning time, making them suitable for detecting emotions in interactive and adaptive systems in real-time. Thus, the field of emotion detection using ML approaches has gained special interest since the early 2000s. Since then, several approaches have been proposed. With this increased attention to this domain, it becomes essential to investigate it and provide a comprehensive systematic review of its trends. Therefore, this paper aims to fill this need by investigating the current state, challenges, and future direction in the area of emotion detection and sentiment analysis. In particular, the paper focuses on machine learning-based approaches for emotion detection and SA. The main goal of the study is to answer the following research question:

RQ1: What are the Machine Learning approaches used for emotion detection and sentiment analysis?

RQ2: Which machine learning models are the most common for emotion detection?

RQ3: What are the best-performing machine learning approaches for emotion detection?

RQ4: What are the data sources used for emotion detection and elicitation?

RQ5: What are the challenges, gaps, and opportunities for future research?

To answer these questions, we conducted a systematic literature review of 123 articles. We formulated the search query based on the research questions, and we ran it on the most common search engines and digital libraries. We carefully reviewed the selected articles and coded them based on several themes, including ML approach and algorithms, data source, domain, datasets, evaluation metrics, and performance. The results reveal that 1) the literature witnesses increasing attention toward ML-based emotion detection. 2) supervised ML approaches are the most common. 3) English text is the most popular data source (or modality) used. And 4) Traditional ML algorithms, such as Support Vector Machine (SVM) and Naïve Bayes (NB), still gain great attention, and they are performing very well. It is worth mentioning that this does not mean necessarily that ML is outperforming. deep learning algorithms. Deep learning approaches have started gaining increased attention recently, and we expect that more deep learning approaches will be introduced (More discussion in Section 5). Therefore, further studies will be needed in the coming years.

Based on the results and the review, the paper highlights challenges and issues that can be summarised as follows: 1) there is a lack of text-based datasets that consider non-English language. 2) Experiments from different studies are not replicated due to the lack of standardised datasets and a unified evaluation procedure. 3) There is limited attention to data sources other than text (e.g. biosignals and body gestures) caused by the lack of available datasets. And 4) many studies focus on evaluating the *Accuracy* of models while neglecting other important factors, such as processing and learning time, which are very important for applying these models for interactive systems evaluation in real-time. Accordingly, we highlight opportunities for future work to advance the state-of-the-art in this area, which helps advance the domain of interactive systems' design and allows for developing interactive systems that can interact emotionally.

2. Background and related work

This section introduces the background concepts and the preliminaries of this work. Specifically, it introduces the concepts of Sentiment Analysis and Emotion

detection and shows their differences. It also provides a brief introduction to ML techniques and defines the scope of this work.

2.1. Sentiment analysis

Sentiment Analysis (SA) is a subclass of affective computing that aims to identify, analyze, and classify subjective information, such as emotions, from the source material. It is also known as Opinion Mining. Although it has gained more popularity in the last decade, SA is dated back to the late 1990s (Tang, Tan, and Cheng 2009). It has a wide range of applications, but it is mainly used to analyze data from web 2.0 resources, such as opinions on social media and reviews on eCommerce and other resources. Such analysis is helpful for several domains, such as politics, education, economics, etc.

Individuals may express sentiments explicitly or implicitly (Saad 2014). For instance, a sentence like 'bad quality pictures' expresses a negative sentiment explicitly, while the sentence 'the battery lasts for 30 min' also expresses a negative sentiment but implicitly. Sentiment analysis involves different tasks, which include subjectivity detection, polarity detection, and sentiment strength detection (Saad 2014). Subjectivity detection is the task of distinguishing subjective sentences (e.g. 'the camera quality is amazing') from other objective sentences (e.g. 'last month I bought a camera'). Subjective sentences are retained and used for analysis as they bear sentiment, while objective sentences are discarded as they express facts. In the Polarity Detection task, the subjective sentences are classified based on the considered sentiment classes (i.e. positive, negative, neutral, etc.). However, in some cases, the sentences cannot be classified as positive or negative sentences. For instance, the sentence 'although it is heavy, the camera has many professional features' holds a negative and a positive opinion. In such cases, the third task (i.e. sentiment strength detection) is utilised to classify sentences into more fine-grained classes (e.g. 'strong positive,' 'positive,' 'week positive,' etc.) (Saad 2014).

2.2. Emotion detection

Affective science has widely studied emotion detection and classification, and several theories and models have been proposed accordingly. According to Imran et al. (Imran et al. 2020), emotion classification models are divided into two categories: discrete and dimensional models. In the discrete models, as the name indicates, emotions are recognised in discrete classes. For instance, the model introduced by Plutchik (Plutchik 1980) concluded eight emotions (anger, fear, sadness, disgust, surprise, anticipation, trust, and joy) and proposed Plutchik's Wheel of Emotions. Another example is the well-known model of Ekman (Ekman 2008), which comprises six emotions (anger, disgust, fear, happiness, sadness, and surprise). These six emotions are known as the basic emotions, and they are thought to be recognisable across cultures (Imran et al. 2020).

The second category is called dimensional emotions. It groups emotions into dimensions instead of discrete emotions. A dimensional theory of emotion can be defined as the theory that classifies emotions by dimensional affective space (Lee, Teng, and Hsiao 2012). Arousal, Valence, and Intensity are examples of emotion dimensions. Although most of the early research considers at least three dimensions, many recent researchers focus on two-dimensional models (Fontaine et al. 2016). There is still no consensus on the optimal number of dimensions to represent the emotion domain (Fontaine et al. 2016). Examples of the dimensional emotion category include the Positive Affect Negative Affect Scale (PANAS) (Watson, Clark, and Tellegen 1988), the Circumplex model (Russell 1980), and the Pleasure, Arousal and Dominance (PAD) model (Mehrabian 1996).

2.3. Sentiment analysis vs. emotion detection

Many researchers use the terms 'Sentiment' and 'emotion' interchangeably because they are both related to human subjectivity. However, the two terms hold slightly different meanings (Chen et al. 2021). Since this study considers both sentiment analysis and emotion detection, this subsection summarises these slight differences.

According to Munezero et al. (Munezero et al. 2014), Sentiment is defined as 'an attitude, thought, or judgement prompted by a feeling.' Emotion, on the other hand, 'refers to a conscious mental reaction subjectively experienced as strong feelings.' The main difference between these two is the duration in which they are experienced (Munezero et al. 2014). That is, sentiments last for a longer period, and they are more stable than emotions (Jain and Kaushal 2018). Also, unlike sentiment, emotions are not necessarily targeted toward an object (e.g. a person, a place, a thing, etc.). For instance, people may wake up feeling happy for no specific reason. Finally, emotion is more sophisticated compared to sentiments.

Despite the differences mentioned above, sentiment analysis (SA), Opinion Mining (OM), and emotion detection and classification are sometimes used

interchangeably, especially by non-psychologists. According to Kaur and Saini (2014), sentiment analysis and opinion mining express a mutual meaning. Opinion mining concerns '*gathering feelings or emotions associated with the text*' or '*extracting the opinions from text*,' and Sentiment Analysis is '*classifying the text according to the sentimental information associated with the text*.' Therefore, Sentiment Analysis and Opinion Mining can be used interchangeably as they represent the same field (Kaur and Saini 2014; Bing and Lei 2013; Medhat, Hassan, and Korashy 2014).

On the other hand, Emotion Detection is slightly different. The main difference between sentiment analysis and emotion detection is that SA uses two or three classes (e.g. negative, positive, and neutral) to classify texts, while emotion detection uses a wider range of classes (or emotions) like joy, fear, anger, brief, surprise or disgust (Kaur and Saini 2014). That is, emotion detection digs deeper into understanding people's feelings, and it provides more fine-grained analysis. Having said that, we can say that sentiment analysis and emotion detection are interrelated domains, such that SA provides an overall view or polarity of opinions, while emotion detection is a deeper and more fine-grained analysis of people's opinions. For instance, in sentiment analysis, the statements '*the item was not as expected*' and '*I hate this item*' are both negative sentences, although there is a significant emotional difference between both sentences. In addition to these theoretical differences, there are technical differences, such that, since emotions are more sophisticated, as mentioned above, their detection and classification are more challenging (Chen et al. 2021).

2.4. Sentiment analysis approaches

Sentiment Analysis approaches can be divided into two main categories (Yiran and Srivastava 2019) and (Anjaria and Reddy Guddeti 2014), Lexicon-based (or lexical) approaches and Artificial Intelligent (AI) or Machine Learning-based (ML-based) approaches. Other researchers (Weichselbraun, Gindl, and Scharl 2010) and (Rohini, Thomas, and Latha 2017) consider the combination of these two categories as a third category, and they call it the Hybrid approach. Lexical approaches rely on *sentiment lexicon* (a.k.a opinion lexicon or tagged dictionaries), which is a collection of sentimental words, such as 'excellent' or 'bad' (Weichselbraun, Gindl, and Scharl 2010). Lexical approaches compare sentiment words with seed words. The sentiment lexicon involves a list of terms. Each term is assigned a value called sentiment value, which is most commonly a numerical value. For instance, the term 'excellent' would have the value (1) as an indication of positive sentiment. On the other hand, a negative term, such as 'terror,' would have the value (−1) assigned to it. The overall score of a document can be assigned one of the multiple categories, usually *positive*, *negative,* or *neutral*. Lexicon-based approaches are divided further into two sub-categories: Corpus-based and Dictionary-based (Rohini, Thomas, and Latha 2017). Since this topic is not the main focus of this paper, we will not go deeper into this discussion.

The other category of SA is the machine learning-based approach. Machine Learning (ML) is a system that learns from prior samples or experience (Rohini, Thomas, and Latha 2017). ML-based SA approaches rely on learning algorithms to understand the sentiments. They use train datasets to train the model on recognising sentiments and classifying emotions. The success of this approach depends, to a high extent, on the algorithm applied and the dataset used. Machine learning techniques are broadly divided into shallow and deep learning (Janiesch, Zschech, and Heinrich 2021), where shallow learning is further divided into supervised, unsupervised, and semi-supervised learning. In supervised learning, the model is trained on a labelled or tagged dataset. On the other side, unsupervised learning models deploy ML algorithms to analyze and categorise using unlabelled data. Section 2 discusses these approaches in more detail.

To summarise, Lexicon-based and ML-based approaches are the two main categories of sentiment analysis. Lexicon-based approaches are easy to perform and receive more attention because of their simplicity (Han et al. 2020). However, the lexicon-based approach is insufficient for large datasets with many aspects or dimensions (Goodfellow, Bengio, and Aaron 2016). On the other hand, ML-based approaches are more flexible, but they are highly dependent on the quality of data (Goodfellow, Bengio, and Aaron 2016). Figure 1 summarises sentiment analysis approaches.

2.5. Machine learning approaches

Machine Learning (ML) is a branch of computer science that imitates the human learning process (Naqa and Murphy 2015). It is part of the Artificial Intelligence (AI) domain, which comprises any technique that helps computers mimic human behaviour and intelligence to solve complex tasks with minimal (or no) human intervention(Russell and Norvig 2020). Machine learning indicates a computer's capacity to improve automatically (i.e. without explicitly being programmed) through experience (Jordan and Mitchell 2015). According to Janiesch et al. (Janiesch, Zschech,

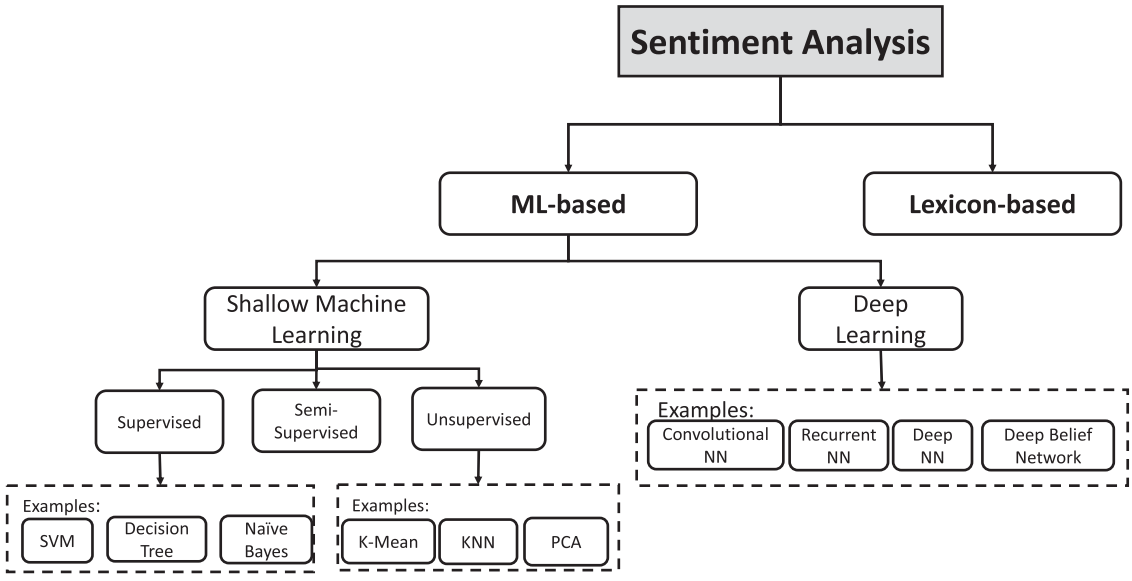

Figure 1. Sentiment Analysis Approaches.

and Heinrich 2021), ML approaches can be broadly divided into two main categories, *Shallow machine learning* and *deep learning*. Shallow (or conventional) machine learning can also be divided into several categories based on the given problem and the available data. It comprises several categories, Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning.

Supervised ML algorithms have shown efficiency in several applications, such as spam detection, weather forecasting, and pricing prediction. They are the most commonly used algorithms. Naïve Bayes, SVM, and Linear Regression are examples of the wide array of supervised ML algorithms. Despite their popularity, these algorithms are associated with several limitations. First, they cannot be applied to all domains and all datasets. Particularly, supervised learning cannot be applied if the data is unlabelled. Second, these approaches are often task-specific, which means that they cannot be generalised for other tasks (Sarkar and Etemad 2020). Third, supervised learning algorithms require a large, annotated dataset to work efficiently. To overcome the limitations of supervised learning, unsupervised ML algorithms are used. These algorithms discover unclear (or hidden) patterns without human intervention. Unsupervised ML uses algorithms to analyze and cluster unlabelled data. Due to their ability to discover similarities between items, unsupervised algorithms became a good solution for many applications, such as customer segmentation, image recognition, and dimensionality reduction. Examples of these algorithms include the Principal Component Analysis (PCA), K-mean clustering, and Singular Value Decomposition (SVD). Nonetheless, unsupervised learning also has some limitations. For instance, it is computationally complex because it requires large amounts of data to analyze and give the outcomes. In between the supervised and unsupervised approaches, there is a third, less common category called Semi-supervised learning. As the name indicates, semi-supervised learning is an intermediate solution between supervised and unsupervised learning. These approaches are used when the data comprises a small portion of labelled data and a large portion of unlabelled data. It uses the small, labelled data to guide classification and feature extraction from a larger unlabelled dataset (Bilgin and Şentürk 2017).

The second category of ML approaches, Deep Learning, is an AI concept modelled on the Artificial Neural Networks (ANN) and neural pathways of the human brain. ANN is typically organised into networks with different layers: an input layer receives the data input, an output layer produces the ultimate result and zero or more hidden layers responsible for learning the mapping between input and output (Goodfellow, Bengio, and Aaron 2016). '*Deep*' refers to the multiple layers between the input and output layers. Deep neural networks have capabilities that allow them to automatically discover a representation of raw data (Janiesch, Zschech, and Heinrich 2021). Common deep learning techniques include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNN), and Deep Belief Networks. Deep learning typically needs less ongoing human intervention. However, it is more complex and requires more computing power. Therefore, deep learning systems require far more powerful hardware and resources, but shallow machine learning

runs on conventional computers. In terms of time, Deep learning systems take more time to set up compared to shallow machine learning systems, which can be set up and operate quickly. Besides, deep learning can accommodate a large volume of unstructured data. Finally, because of deep learning characteristics, it is more common in complex and autonomous applications, such as robots and self-driving cars. DL is particularly useful in domains with large and high-dimensional data. However, for low-dimensional data input, shallow ML can outperform deep learning, especially in cases of limited training data availability (Zhang and Ling 2018; Rudin 2019).

2.6. Study scope

As mentioned before, emotion and emotion detection on one side and sentiment and sentiment analysis on the other side have some differences. However, since many researchers use these aspects mutually or interchangeably, this study considers works done in both domains. Regarding the sentiment analysis approaches, the study considers ML-based approaches while it excludes other approaches (namely, lexicon-based approaches). Among ML-based approaches, this study considers all retrieved papers that talk about ML algorithms, whether shallow or deep learning. It is worth mentioning that the sentiment analysis domain, and AI-based sentiment analysis, in particular, is huge. Sentiment analysis includes several aspects, such as shallow ML, deep learning, affective computing, symbolic and asymbolic AI, and explainable AI for sentiment analysis. Having such a multidimensional domain makes it difficult to comprehend this domain in a single study. With respect to emotion categories, this paper considers both discrete and dimensional emotions. That is, it includes papers that focus on both categories. To summarise, this paper surveys research that proposes an ML-based approach for sentiment analysis or emotion detection.

2.7. Related work

Various studies have been conducted to compare different ML techniques for sentiment analysis and emotion detection. For example, Narendra et al. (Narendra et al. 2016) conducted a comparative study focusing on user reviews of movies. This paper present a comparative study of sentiment analysis of movie reviews using different machine learning techniques, but it does not provide a summary of previous studies. A similar study is provided by Nabil et al., but this paper focuses on users' reviews on Amazon. This paper discusses the different sentiment analysis approaches and compares the performance of the different machine learning algorithms. More comparative studies were conducted to compare the performance of different ML techniques in analyzing sentiment and detecting emotions in different domains, including Gujarati film audits (Shah and Swaminarayan 2022), Social Media posts (Hammad and Al-Awadi 2016; Singh, Verma, and Kumar 2020), and News (Vaseeharan and Aponso 2020). All these studies and more compare different ML techniques for sentiment analysis, but they did not explore and provide a comprehensive view of state of the art in the domain of ML techniques for sentiment analysis and emotion detection.

Other researchers have conducted systematic reviews to explore various aspects in the domain of ML for sentiment analysis and emotion detection. For instance, Machado et al., (Machado, Ribeiro, and e Sá 2019) conducted a systematic literature review that aims to report the state-of-the-art machine learning techniques for sentiment analysis applied to texts of reviews, comments and evaluations of scientific papers. The paper provided a summary of the publications' trends per year and explored sentiment analysis techniques. However, this paper was limited to text-based sentiment analysis and lacked a discussion about several Ml-related aspects.

Bansal and Singh (Bansal and Singh 2016) surveyed 32 articles on using machine learning approach for sentiment analysis, considering various dimensions like granularity, type of data used, polarity, and algorithms. The paper considered existing literature in English and Chinese language only. Also, this paper is relatively old as it includes papers published between 2011 and 2016.

Although most of the available work and reviews focus on the English language, some reviews were conducted in other languages. For instance, Sagnika et. al., (Sagnika et al. 2020) explored studies on the sentiment analysis methods applied to languages other than English. The authors provided details on the tools used, domains, pros and cons, the efficiency of the methods covered, and the associated challenges. The paper involved 49 articles covering methods that analyze translated data and methods that analyze available data in the target language. Although the paper provided details about the approaches used, it was limited to languages other than English and focused on text-based sentiment analysis.

Some research articles targeted particular domains. For instance, sentiment analysis in the education domain was explored in different studies, such as (Han et al. 2020; Kastrati et al. 2021; Mite-Baidal et al. 2018). Kastraki et al., (Kastrati et al. 2021) and Kastrati

et. al., (Mite-Baidal et al. 2018) systematically reviewed articles in the education domain. In (Kastrati et al. 2021), the authors extended the review presented in (Mite-Baidal et al. 2018), and they reviewed 92 articles published between 2015 and 2020. They classified the research and results of the application of machine learning solutions for sentiment analysis in the education domain. Another domain-specific study was conducted by Jain and Kaushal (Jain, Pamula, and Srivastava 2021). The study reviewed 68 papers on the use of machine learning techniques for consumer sentiment analysis in online reviews. This study provided information about ML applications, datasets, and other descriptive analyses, and it found that ML has a strong potential in the hospitality and tourism domain.

This section discusses the most recent and related reviews in sentiment analysis and emotion detection approaches. As discussed above, various reviews and comparative studies have been conducted. Although these reviews discuss different aspects related to the domain of our paper, they suffer from one or more of the following limitations: 1) the study is limited to a single domain (Han et al. 2020; Jain and Kaushal 2018; Kastrati et al. 2021; Mite-Baidal et al. 2018), 2) the study considers a relatively limited number of papers (Bansal and Singh 2016; Jain and Kaushal 2018; Sagnika et al. 2020), 3) the study is limited to a particular language (Bansal and Singh 2016; Sagnika et al. 2020), and 4) most of the reviews explore ML approaches for sentiment analysis based on text while neglecting other modalities (e.g. body gestures or voice recognition). Therefore, it becomes necessary to conduct a study that addresses these limitations. Also, the noticeably increased interest in this domain necessitates a continuous exploration and investigation of the domain to give researchers a clear and accurate overview of the current state of the art. To fill these needs, we conducted this literature review.

3. Methods and material (Study design)

The main goal of conducting this review is to systematically analyze and provide a detailed summary of the current literature to answer the research questions. This paper presents a systematic literature review. The paper followed a well-established procedure for conducting a systematic literature review by Kitchenham (Kitchenham 2004). Specifically, this study went through three main stages: searching, evaluating, and synthesising papers in the area. In the first stage, we searched the targeted electronic databases using some search queries. Then, we evaluated the retrieved papers against the inclusion/exclusion criteria to retain only related papers. Finally, we synthesised by studying the approaches deployed in the papers, trends in the domain, challenges, and outcomes.

3.1. Search method

A search of key terms related to emotion detection and SA was conducted using selected databases. The search was restricted to English-language publications that concern using ML approaches to analyze and classify people's emotions. Search results were limited to peer-reviewed articles published in the last two decades. This time interval has been chosen because it is the time when sentiment analysis and emotion detection domain gained wide attention. Older studies are less relevant today, and two decades is considered appropriate to capture interesting trends. Three data sources were considered to locate the relevant articles. These data sources are ACM, IEEE Explorer, and Scopus digital libraries. These are the popular database for research in this area. We used an automatic search method such that we ran the search query on the selected databases instead of searching several venues manually. Based on the research question, we extracted the main keywords, which include the terms: 'Emotion,' 'Sentiment,' 'Machine Learning,' and 'detection,' along with their synonyms. Then we formulated the search query based on these keywords and their synonyms using both 'OR' and 'AND' conjunctions; see Figure 2 for example, which depicts the main search query.

3.2. Selection criteria

After running the above search query, we selected the most related papers. We first filtered the papers based on the title and abstract review. Then, we evaluated the selected papers against the inclusion/exclusion criteria and filtered them based on the full-text review. – Table 1 summarises the inclusion and exclusion criteria. As the table shows, we included the full peer-reviewed papers that talk about using ML for SA or emotion detection. Any paper written before the last two decades, not in English or not peer-reviewed was excluded. Also, literature review and survey papers were excluded. Figure 3 depicts the paper selection process. After removing duplicated papers and after applying the inclusion and exclusion criteria, we retained 120 papers. Following the snowballing approach (i.e. checking the references of the included papers), we could include three more related papers. Finally, we retained and investigated 123 papers.

("machine learning" OR "ML" OR "Artificial Intelligen*" OR "AI" OR "*learning" OR "Classifi*")
AND
("emotion" OR "mood" OR "affect" OR "sentiment" OR "feel*" OR "opinion")
AND
("Prediction" OR "elicit*" OR "analysis" OR "detection" OR "recogni*" OR "mining")

Figure 2. Search query.

3.3. Data extraction

The last step in this study is data coding. After applying the inclusion and exclusion criteria and evaluating the papers against them, the full-text version of all included papers was retrieved to start reading and extracting data. The extracted data includes details about each study's title, year, modality, datasets, ML techniques and algorithms, evaluation metrics, and application domain. In addition, we provided a comparison of the most popular ML algorithms based on prior comparative studies. The following section discusses this data and presents the results related to each one. Both authors were involved in the selection and analysis process. One author did the first round of inclusion/exclusion, while the second one did a second evaluation. Any conflict between both researchers' decisions was resolved through a discussion meeting.

4. Results analysis

This section discusses the results obtained from this study. It is divided into several subsections according to the extracted data.

4.1. Descriptive analysis

This subsection provides a general overview of the papers considered in this study. It shows their distribution over time, the literature trend in the SA and emotion detection domains, and the distribution based on the application area or domain.

Table 1. Inclusion and Exclusion Criteria.

Inclusion Criteria	Exclusion Criteria
The paper concerns machine learning for sentiment analysis or emotion detection	Not peer-reviewed. So, reports or proposals were excluded
Evaluate the proposed approach and includes details about the evaluation methodology	Written not in the English language
Written in the English language	Published before 2001
Full peer-reviewed research paper	Literature review, survey, non peer-reviewed articles

4.1.1. Time-based analysis

Figure 4 depicts the number of publications per year through the considered time interval (i.e. from 2001 until 2020). As the figure shows, the number of papers published each year seems to have increased steadily from 2010 and has increased deponently since 2017. Besides, the highest number of publications were found in the last three years. This noticeable increase in publications demonstrates the increasing interest in the SA and emotion detection domain in recent years.

To get a clearer picture of the publication trend, Figure 5 shows the percentage of publications in every quarter of the last two decades. It divides the time into four quarters, as follows: Q1 (2001–2005), Q2 (2006–2010), Q3 (2011–2015), and Q4 (2016–2020). It is clear from the figure that Q4 witnesses the highest number of publications. Specifically, it has more than five times the number of papers published in the third quarter. Moreover, the figure also shows that the number of publications is increased in Q2 compared to Q1. However, the increase is marginal compared to the difference between Q3 and Q4.

4.1.2. Publication type

As mentioned above, we only included peer-reviewed papers. This subsection shows the results regarding

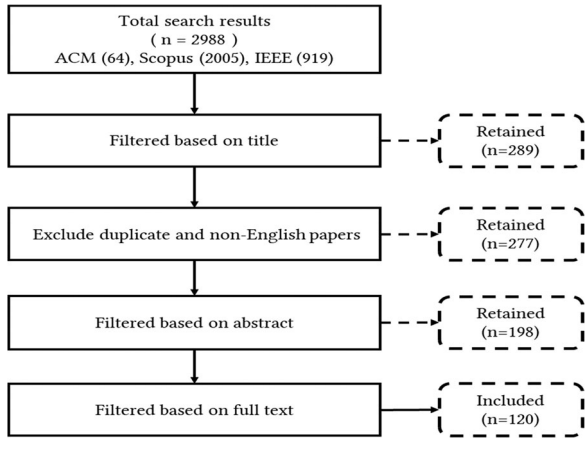

Figure 3. Selection Process.

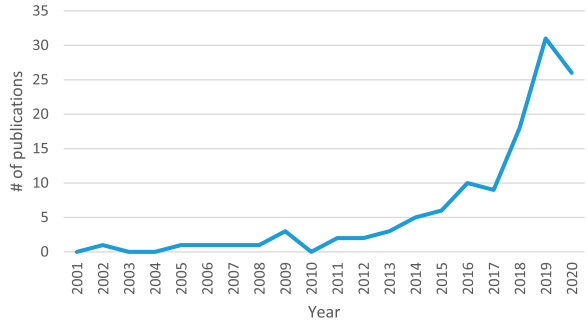

Figure 4. Number of publications per year.

the publication type (i.e. whether it is a conference or journal paper). Figure 6 shows the distribution of the papers based on their types. As the figure depicts, most papers (81%) are conferences papers, while (19%) are journal papers. We also noticed that the publication venues span various areas, including but not limited to Engineering, Sustainability, Communication, the Internet of Things, Power and Energy, and Education. Table 2 depicts the most popular venues. Due to the wide variety of venues, and for clarity purposes, the table does not list all venues.

4.1.3. Application area

As mentioned before, SA and emotion detection is gaining special popularity in a wide range of domains. In this study, we investigated these domains and discovered how popular SA and emotion detection are in each domain.

Our study shows that the work in this domain is spread over more than fifteen areas. These areas include eCommerce (e.g. (Ghosh and Sanyal 2018; Goldberg and Zhu 2006; Goodfellow, Bengio, and Aaron 2016; Gupta et al. 2018; Hamdan, Vigier, and Wantiez 2017; Hammad and Al-Awadi 2016; Han et al. 2020; He et al. 2018; Hu et al. 2013; Van Huynh et al. 2019; Imran et al. 2020; Indulkar and Patil 2021; Ismail et al. 2018; Jain and Kaushal 2018; Jain, Pamula, and Srivastava 2021; Jang et al. 2014; Janiesch, Zschech, and

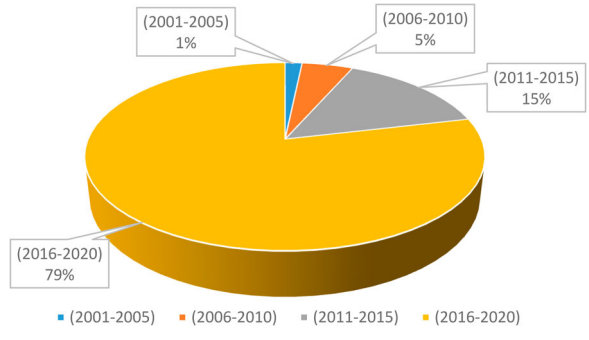

Figure 5. Distribution of publications.

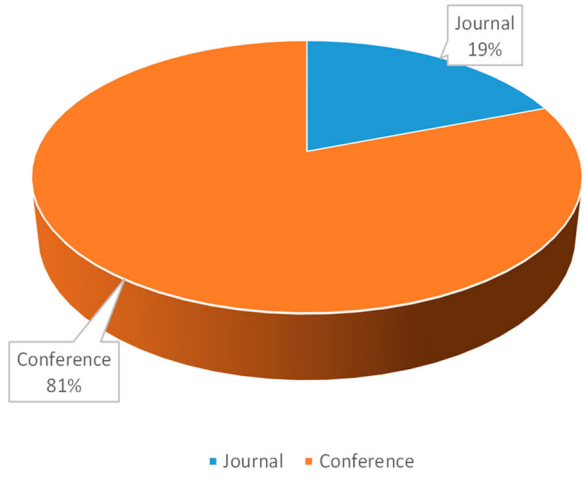

Figure 6. Publication Distribution based on venues.

Heinrich 2021; Javaid et al. 2021; Jordan and Mitchell 2015; Joseph, Pramod, and Nair 2018; Kannadaguli and Bhat 2018; Karim and Das 2018; Kastrati et al. 2021; Kaur and Saini 2014; Khan et al. 2020; Kitchenham 2004; Kukolja et al. 2014; Lalata et al. 2019; Lee and Sdn Bhd 2011; Lee, Teng, and Hsiao 2012; Lin and He 2009; López and Cuadrado-Gallego 2019; Ly et al. 2018; Machado, Ribeiro, and e Sá 2019; Majeed, Mujtaba, and Beg 2020; Malheiro et al. 2013; Mankar et al. 2018; Medhat, Hassan, and Korashy 2014; Mehrabian 1996; Mite-Baidal et al. 2018; Muhammad and Shamim Hossain 2021; Muhammad et al. 2020)), microblogs (e.g. (Bilgin and Şentürk 2017; Elbagir and Yang 2018; Qian, Niu, and Shi 2018; Zhang, Lu, and Song 2017; Ismail et al. 2018; Anjaria and Reddy Guddeti 2014)), multimedia (e.g. (Appel et al. 2016; Chen et al. 2019; Tripathy, Agrawal, and Rath 2016; Van Huynh et al. 2019; Ly et al. 2018)), learning (e.g. (Lalata et al. 2019; Alm, Roth, and Sproat 2005; Ashwin et al. 2016; Pong-Inwong and Kaewmak 2017; Sultana et al. 2018)), finance (e.g. (Chiong et al. 2018; Mankar et al. 2018)), technology (e.g. (Joseph, Pramod, and Nair 2018; Muhammad and Shamim Hossain 2021; Studiawan, Sohel, and Payne 2020; Khan et al. 2020; Alkalbani et al. 2016)), and more. The cluster bar chart in Figure 7 shows the domains and depicts the distribution of the

Table 2. Most popular venues.

Venue	Type	Publisher
International conference on machine learning and computing	Conference	ACM
Advances in Intelligent Systems and Computing	Journal	Springer
Expert Systems with Applications	Journal	Elsevier
International Conference on Computer and Information Technology	Conference	IEEE
International Conference on Electronics Computer and Computation	Conference	IEEE

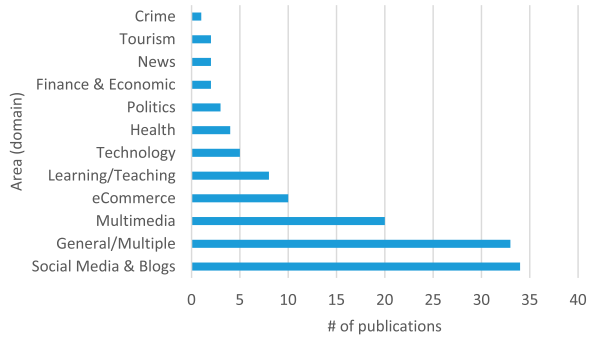

Figure 7. Distribution based on domain.

publications over these domains. The figure shows that SA and emotion detection on posts on social media and blogs gained special interest among other domains. The reason behind this popularity is the rise in the popularity of social media in the last decade and its use for diverse purposes, including health, marketing, politics, and even education. Social media plays an essential role in understanding users' attitudes and feelings. Also, social media and microblogs contain a massive amount of data about various topics. This data, in most cases, are available to the public. Multimedia and eCommerce are the second and third most popular areas. Multimedia combines the domains of movies and videos, music, and books. Emotion detection is used in these domains to understand users' opinions and feelings toward products (Ghosh and Sanyal 2018). More specifically, the studies in these domains mainly used datasets of customers' or users' reviews.

Some papers mentioned the domain, neither explicitly nor implicitly. These papers are categorised as 'General/Multiple.' It is worth noting that even those papers that mention a specific application area are not necessarily limited to the mentioned area. However, in some cases, the application areas are considered as use cases to apply the proposed approach.

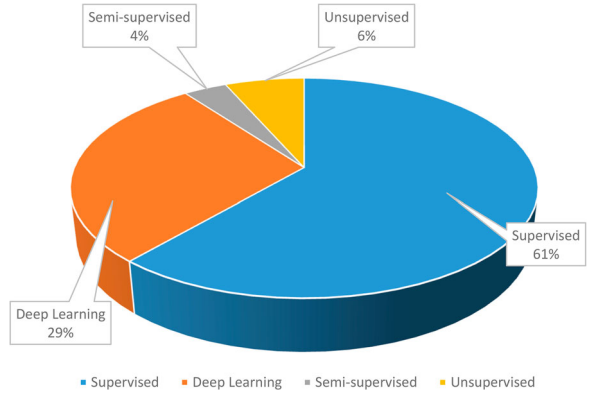

Figure 8. Distribution based on ML approaches.

4.2. Learning approaches

As mentioned in Section 2, ML approaches can be divided into three categories, supervised, unsupervised, and semi-supervised learning approaches. This section discusses our findings related to these categories and their different ML algorithms.

Figure 8 shows the percentages of deploying each ML approach in the considered articles. It clearly shows that the supervised ML approach (e.g. (Appel et al. 2016; Boiy and Marie-Francine 2009; Anjaria and Reddy Guddeti 2014; Lee and Sdn Bhd 2011; Tripathy, Agrawal, and Rath 2016; Hamdan, Vigier, and Wantiez 2017; Majeed, Mujtaba, and Beg 2020; Alm, Roth, and Sproat 2005)) is the most common one. This popularity can be explained by the simplicity of implementing supervised algorithms and the increasing availability of labelled data or data labelling techniques, especially for text-based analysis. As mentioned above, social media and online blogs are primary sources of information for many parties because they hold a huge amount of data. However, the data provided in such resources are usually unlabelled. Thus, researchers have to label this data manually (crowd-based) or automatically (Roh, Heo, and Whang 2021). Manual labelling can be done in-house (by the research team), using crowdsourcing, or by third parties (outsourced). However, manual labelling is time-consuming and ineffective, especially for large datasets. On the other hand, automatic labelling, which is done using data programming, is faster and less costly.

Our results show that Deep learning emerged as the second most commonly considered approach. Specifically, (29%) of the included papers considered at least one deep learning algorithm. The popularity of deep learning is expected due to its advanced capabilities and characteristics, which allow building models for complex domains and large unstructured data. Finally, our results show that unsupervised learning (e.g. (Lin and He 2009; Pitogo, Diane, and Ramos 2021; Hu et al. 2013; Xu et al. 2019; Pong-Inwong and Kaewmak 2017; Neumann and Vu 2019; Karim and Das 2018)), and semi-supervised learning (e.g. (Bilgin and Şentürk 2017; Gupta et al. 2018), and (Goldberg and Zhu 2006)) are the least considered approaches. It is worth mentioning that some papers (e.g. (Karim and Das 2018)) used more than one algorithm belonging to different categories (e.g. supervised and unsupervised). These papers are counted for each category.

4.2.1. Machine learning algorithms

Hundreds of ML algorithms have been introduced to the ML literature. Different algorithms can be applied

to different domains. In regard to SA and emotion detection, our study found more than twenty (20) algorithms were used by the papers considered in this study. Figure 9 depicts the popularity of each ML algorithm. Specifically, it shows the most commonly used algorithms based on the considered papers. As the figure shows, SVM is the most commonly used algorithm. It is used by 54 papers (44%) of the considered sample. Naïve Bayes (NB) comes next in popularity (40 papers), followed by Logistic Regression, Convolutional Neural Network, and Random Forest. On the other hand, Gaussian Process Regression, Gaussian NB, and Linear Regression are the least common algorithms.

An interesting observation here is that the conventional ML algorithms still capture researchers' attention in the SA and emotion detection domains. These algorithms still show promising results (as we will see in the next subsection). More interestingly, the focus on these algorithms is not diminished and is still increasing. For instance, among the 39 papers that deployed NB, 28 papers were published in the last five years.

Ensemble learning is the process of combining multiple learning algorithms (Pong-Inwong and Kaewmak 2017). It shows efficiency in improving the learning performance of the models. Ensemble learning is found in nine (9) papers only. This is because it is relatively new compared to other ML algorithms. Nonetheless, our review shows that all nine papers using Ensemble learning were published in the last decade, and seven out of the nine papers were published in the last five years. This observation indicates increasing attention toward this learning approach.

4.2.2. Comparative analysis

The data and results presented in Section 4.2.1 summarise the popularity of different ML algorithms. To give more insights in this direction, we investigated the literature for comparative analysis. We found that several

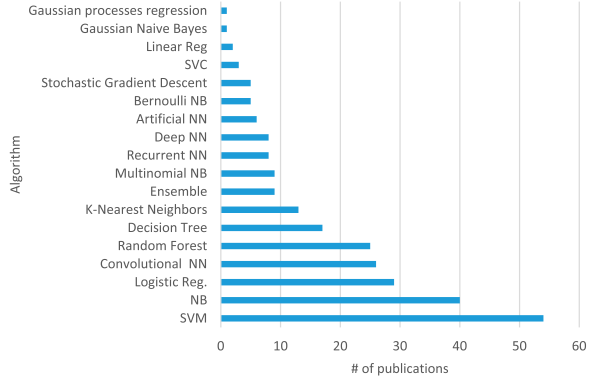

Figure 9. Distribution based on ML algorithms.

comparative studies have been conducted to evaluate the performance of different ML algorithms for emotion detection. Table 3 summarises the results of sixteen studies. It shows the metrics used for evaluation along with the outcomes. The 'Outcome' column represents the conclusion of these papers regarding algorithms' performance in terms of the considered metric. The algorithms are listed in descending order (from the most efficient to the least). As the table shows, most of the papers use *Accuracy* as the main evaluation metric, and some of them use the *F1-measure*.

Table 3 also shows that most of these studies consider at least one of the three most common ML algorithms (SVM, NB, Logistic Regression). Specifically, SVM was considered by (80%) of the papers, NB was considered by (56%), and Logistic Regression was considered by (44%) of the papers (as depicted in Figure 10). This confirms the popularity of these algorithms. Out of these studies, twelve concluded that SVM (the most common algorithm) is either the first, second, or third most efficient one. Specifically, four studies found it as the most efficient algorithm, six studies found it the second most efficient, and two found it the third most efficient algorithm. The second most common algorithms (NB) have been seen as the first, second, and third most efficient algorithm by three, four, and four studies, respectively. Finally, one paper found Logistic

Table 3. Summary of comparative studies.

Reference	Metric	Outcome (Descending order)
(Rahman et al. 2019)	Accuracy	NB, RT classification, DT classification, SVM, LDA, Log. Reg, ANN, SVR, PCA, KNN.
(Sharma and Dey 2012)	Accuracy	SVM, NB, Decision Tree, Max Entropy (logistic reg), KNN
(Saad 2014)	Precision, Recall, F1-measure	BMNB, SVM, NB, MNB, J48
(Poornima and Sathiya Priya 2020)	Accuracy	Log. Reg., SVM, MNB
(Becker, Moreira, and dos Santos 2017)	F1-measure	RBF, SMO, NB
(Indulkar and Patil 2021)	Accuracy	Random Forest, Log. Reg, MNB
(Sajib, Shargo, and Hossain 2019)	Accuracy	SVM, NB, Log. Reg
(Zhang and Zheng 2017)	Accuracy	ELM, SVM
(Malheiro et al. 2013)	F1-measure	SVM, NB, C4.5, KNN
(Doma and Pirouz 2020)	Accuracy	KNN, SVM, Decision trees, Logistic Regression, and LDA
(Vijayakumar, Flynn, and Murray 2020)	Accuracy	NN, SVM, KNN
(Vijayakumar, Flynn, and Murray 2020)	Accuracy	SVM, LR, LDA
(Vaish and Kumari 2014)	Accuracy	MLP, C4.5, Decision Table. NB
(Kukolja et al. 2014)	Accuracy	MLP, NB, SVM, KNN, RIPPER, RF, C4.5
(Nugrahaeni and Mutijarsa 2017)	Accuracy	KNN and Random Forest, SVM

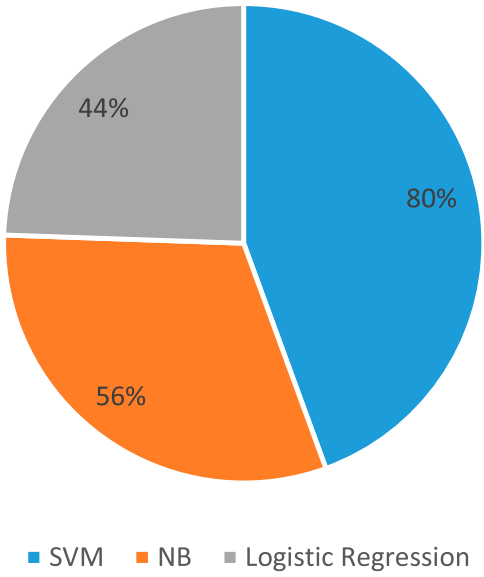

Figure 10. Percentages of comparative papers considered the most common shallow ML algorithm.

Regression (the third most common algorithm) as the most and third most efficient, while two papers found it the second most efficient.

In regard to the comparison between the three most common algorithms, five studies found SVM more efficient than NB, and four studies found it more efficient than Logistic Regression. NP has shown superiority over SVM in two studies and superiority over Logistic Regression in three studies. Finally, Logistic Regression overcomes SVM and NP in one and two studies, respectively. These comparative results indicate that SVM shows superiority over other algorithms in terms of popularity and efficiency. However, this conclusion cannot be generalised because these studies are not benchmarked. For example, the algorithm's performance could change significantly based on the dataset characteristics. In future work, we plan to compare these algorithms in several datasets.

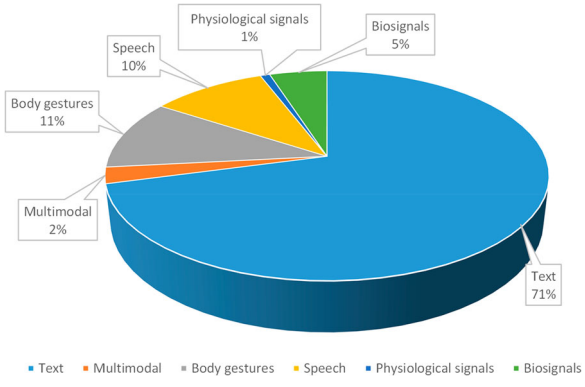

Figure 11. Distribution based on modality.

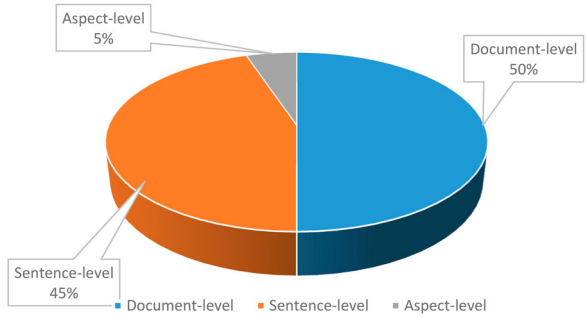

Figure 12. Distribution based on the text level.

4.3. Modality analysis

Emotions and sentiments can be observed and evaluated based on several modalities (data sources or inputs). Our study shows five common modalities, which are text, speech, body gestures (e.g. eye movement, body movements, facial expressions, etc.), physiological signals, and bio-signals (e.g. electrodermal activity, electrocardiogram (ECG), skin temperature, and photoplethysmography). In addition, some papers use a combination of these modalities; We categorise these papers under the *Multimodality* category.

The results presented in Figure 11 show that most of the work done in the last two decades focuses on text-based emotion detection. This popularity can be explained by multiple reasons, including 1) the availability of a large number of labelled and unlabelled text datasets due to advancement in new media, 2) analyzing and classifying text is simpler than other data sources, 3) several algorithms have been validated and used for text analysis, and 4) text is used since the early days of SA and emotion detection. On the other hand, physiological and biological signals are the least common input sources. This does not necessarily mean they are inefficient, but it can be explained by the difficulty and special equipment required to acquire data to use these sources. This equipment is not available for most researchers, which represents a barrier to conducting research that relies on these data sources.

4.3.1. Text dimensions

Text-based SA can be classified into three main levels: document-level, sentence-level, and aspect-level (or feature-level) (Medhat, Hassan, and Korashy 2014): Document-level analysis classifies the document as a whole unit. That is, it considers the whole text as a single document discussing one topic with one sentiment. The sentence-level analysis is a more fine-grained classification. It classifies the sentiment presented in each sentence. To do so, it first classifies sentences into subjective and objective sentences. Subjective sentences are further

analyzed to classify the sentiment, while objective sentences are ignored. Some research (Wilson, Wiebe, and Hoffmann 2005) considers sentence-level analysis similar to the document-level supposing a sentence is just a short document. The third level, the aspect level, is used to obtain more details on the opinions regarding a specific aspect of the entity. An important step in the aspect-level analysis is to identify the important aspects of each entity. For instance, a user may give a review about the sound and the camera of the phone (e.g. the camera is amazing, but the sound is a bit low).

Figure 12 summarises the results regarding text level. It shows that the papers are almost evenly distributed between document and sentence levels, where document-level is applied in 50% of the documents and sentence-level is applied in 45% of the papers. Aspect-level is the least considered dimension. Only 5% of the studies considered this dimension. Many researchers do not consider the Aspect-level because it is more difficult compared to document and sentence-level analysis. The simplest shape of analysis is to classify a whole document as 'positive' or 'negative' (Singh et al. 2013). In addition, there is a lack of aspect-level labelled data, and the available datasets for this task are relatively small (Zhuang and Tieyun 2019). Besides, it is not easy to annotate aspect-level data (Zhuang and Tieyun 2019) and (He et al. 2018).

4.3.2. Language

Another aspect related to text data is language. This aspect is also related to the speech modality. The results show that the reviewed papers considered several languages. Figure 13 summarises these results. Unsurprisingly, we found that English is the most commonly considered language. This is because most of the considered publications are published in English. Also, most of the text resources and datasets re in English. Hindi and Chinese come next to English with six and five papers, respectively. It is worth mentioning that the 'Hindi' category included the standard Indian language and other languages used by some Indian communities, such as the Kannada and the Malayalam languages.

4.4. Data sets

Dataset is crucial for any ML algorithm and supervised learning algorithms, in particular. Dataset size, content, and quality may substantially affect ML performance. This review found that researchers use a wide range of datasets. Some of these datasets are available online, whereas others were collected by the authors. Figure 14 shows the most popular datasets based on our study. Specifically, the figure shows the number of articles that used each dataset in their studies. Our study found more than 60 distinct datasets used by the articles considered in this review. For clarity, Figure 14 only shows datasets that are used by two or more articles.

The 'Self-gathered' category indicates articles where the authors collected data by themselves. It is notable from the figure that 'self-gathered' datasets are used by more than half the articles (65 articles). The massive amount of data available online and the availability of data scrapping tools are key reasons leading authors to self-gather their data. Another reason is the lack of datasets in some domains. For instance, (Soumya and Pramod 2020) collected the data from Twitter because of the unavailability of sentiment-tagged datasets in the Malayalam language (the scope of their study). For the same reason, (Jang et al. 2014) conducted a study to collect tagged data using bio-signals. It is worth noting that most of the self-gathered datasets were collected from eCommerce websites (e.g. Amazon) or social media (e.g. Twitter and Facebook). Figure 14 summarises the most common datasets.

4.5. Evaluation metrics

An essential step for any ML implementation is the evaluation of the model. The evaluation aims to show

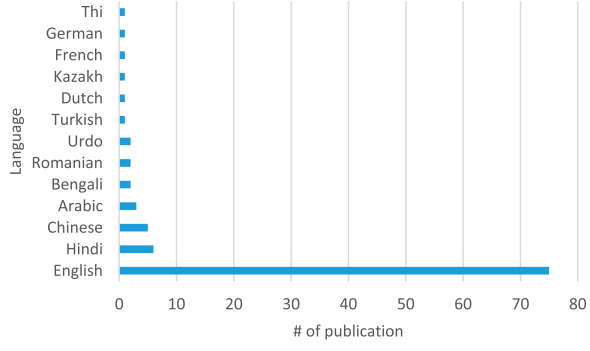

Figure 13. Distribution based on languages considered.

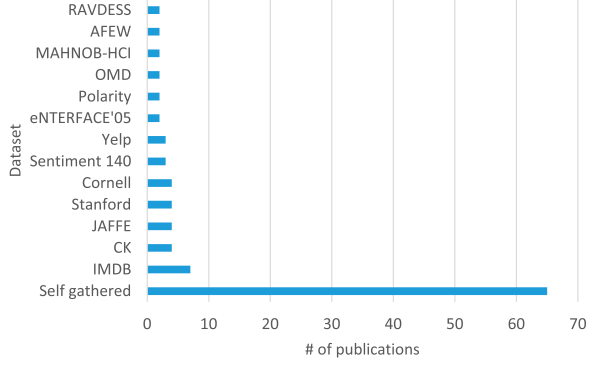

Figure 14. Distribution based on datasets.

how good the ML model is. Several measures are used to evaluate whether or not a proposed learning approach achieved the desired objective. Figure 15 depicts the popularity of the evaluation metrics used in the considered studies, and Table 4 summarises these metrics. *Accuracy*, *Precision*, *Recall*, and *F1-score* were found to be the most popular measures. These measures are commonly used in the ML domain, despite their limitations and biases (as discussed in Section 25). It is worthwhile to mention that our study revealed more than fifteen different measures. However, for clarity purposes, Figure 15 only shows measures used in at least two publications.

5. Discussion and future direction

In the previous Section, we summarised our results regarding using ML for emotion detection and SA. Overall, the results demonstrate increasing work in this domain. Several ML techniques have been deployed, and various algorithms have been implemented. In addition, the results presented in Section 4.1.1 show that the last five years witnessed a noticeable increasing interest in the research related to ML-based emotion detection. This increased attention reflects the importance of this domain theoretically and practically. Nonetheless, the domain has not been fully explored yet. Several aspects have not been investigated yet, and various challenges have not been resolved. This section discusses the findings, highlights the challenges and limitations, and suggests future directions based on the results obtained in this work.

5.1. Machine learning approaches

Section 4.2 presented the ML approaches and algorithms used in the literature. The results show that supervised learning is the dominant approach among the major categories of ML. 61% of the papers deployed at least one supervised learning algorithm. Among these algorithms, SVM, Naïve Bayes, and Random Forest are the most commonly deployed algorithms. This popularity does not necessarily mean supervised learning is the most efficient approach. But it could be because these models are simple, easy to implement, fast, and not costly in terms of time and resources. Thus, they are easier to deploy and to consider in research works. Also, the efficiency of shallow ML algorithms and their low cost makes them a good choice for small and medium-size projects and companies. On the other hand, more advanced approaches, such as deep learning, are more sophisticated and time-consuming. Deep learning algorithms require far more powerful hardware and resources. Thus, deep learning is more common in complex and autonomous applications, such as robots and self-driving cars because of its characteristics.

In addition, we noticed that most shallow ML approaches (81%) are focused on text modality. An expected reason is the availability of text-based datasets and data resources, such as social media and other websites. On the other side, we noticed that deep learning approaches are more included in analyzing modalities other than texts, such as body gestures (24%), speech (24%), and biosignals (12%). This is because deep learning algorithms have characteristics and capabilities that allow more accurate results for unstructured data (Janiesch, Zschech, and Heinrich 2021).

Therefore, given the simplicity, low cost, and high performance of the shallow ML approaches, we recommend using them whenever they are sufficient for the application at hand. However, more advanced algorithms can be used for more complex applications, especially in domains with large and high-dimensional

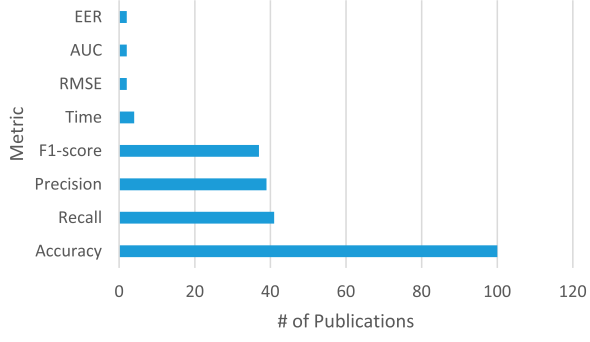

Figure 15. Distribution based on evaluation metrics.

Table 4. Most common metrics (Saad 2014; Powers 2020).

Dataset	Definition
Accuracy	The ratio of correctly labelled (positive or negative) observations to the total observations.
Precision	(a.k.a Confidence) denotes the ratio of correctly labelled positive observations to the total number of observations labelled as positive.
Recall	(a.k.a. Sensitivity) denotes the ratio of correctly labelled positive observations to the actual number of positive observations
F1-score	A measure that combines Precision and Recall
EER	*Emotion Error Rate* (EER) can be defined as the ratio of the number of emotions misclassified to the total number of emotions used for testing (Kannadaguli and Bhat 2018)
AUC	Area Under Cover is evaluated by plotting the True Positive Rate (or Recall) against the False Positive Rate (FPR) (López and Cuadrado-Gallego 2019)
RMSE	*Root Mean Square Error* denotes the differences between predicted and observed (or actual) values (Chai and Draxler 2014).
Time	Represent different times. Some papers consider computation time (Chiong et al. 2018), Execution time (Joseph, Pramod, and Nair 2018), training time, and testing time (Zhang and Zheng 2017).

data. Despite some comparative studies, more comprehensive studies are needed to get more inclusive insights on the best performing algorithms in terms of different measures. This is one important and urgently needed future research direction.

5.2. Data source-related aspects

As mentioned above, the high availability of datasets in many domains leads to higher interest in ML approaches for emotion detection. The selection of a dataset is not straightforward. ML results are very sensitive to the datasets. In addition to the conventional dataset-related issues, such as imbalanced data and overfitting. Other challenges face researchers when it comes to the dataset. First, the abundant number of available datasets makes it difficult for researchers to select the best-fit dataset. Our results found that 123 studies used more than 60 different datasets. Second, the lack of dataset standardisation; the high number of available datasets complicated the standardisation of these datasets. Third, most of the available datasets are text-based (i.e. they consist of texts, such as reviews or tweets). However, as mentioned in Section 4.3, there are other modalities, such as biosignals and face gestures. These modalities are underestimated in the literature, such that a rare number of datasets are available (Song, Lu, and Yan 2020). Due to the limited number of available datasets, these modalities did not gain enough attention in the literature.

As future research directions related to the dataset theme, we invite researchers to collaborate with businesses to build a repository of unified datasets. We also invite them to consider various application areas and the differences between these areas to enhance the repository's quality and increase its benefits for researchers and practitioners. The availability of such unified datasets can serve as a benchmark for evaluating and comparing various ML algorithms for emotion detection. Besides, more attention should be paid to datasets besides text which is the focus of many existing research, such as datasets that involve physiological and biosignals.

5.3. Language aspects

Another aspect related to the data source (text-based data source, in particular) is the language. The results show that English is the dominant language. More than 60% of the papers considered datasets of English content (such as reviews, tweets, or blogs). This indicates a lack of studies that discusses other languages. There are several reasons behind this issue, which includes: 1) the limited number of datasets that consider other languages compared to English language datasets, 2) limited resource in some countries, and 3) tight fund for research in some continents that speaks other languages. Thus, we invite researchers to focus on analyzing emotions in other languages. In addition, the efficiency of the ML algorithms can be compared based on language. That is, the efficiency of an algorithm can be assessed in different languages. Based on these comparisons, a comprehensive study is required to provide insights on the best ML approach for emotion detection based on different languages (i.e. to provide a map between ML algorithms and languages).

5.4. Evaluation-related aspects

This is a very important aspect because selecting one algorithm rather than the other is done based on comparing the evaluation results of several algorithms. This aspect is, to a great extent, related to the data source aspects because the dataset is an essential component of the evaluation process. Results presented in Section 4.4 show that most of the datasets were used by only one of the studies (i.e. each study uses a different dataset), which indicates a lack of unified datasets. This, in turn, complicates the evaluation and comparison process. Also, the results of such evaluations are not replicated for several reasons, including the sketchy details available about the experimental procedure and the use of different datasets and experimental settings. Thus, the literature need unified datasets and benchmark settings for evaluating emotion detection approaches.

- **The issue of a single evaluation metric**. The results presented in Section 4.5 show that Accuracy is the most commonly used metric. Also, some papers only reported Accuracy as a sole performance indicator. Reporting a single metric does not reflect the whole picture of the model's performance. The reality is that the use of a single metric only may lead to wrong conclusions. For instance, Accuracy is inappropriate if the dataset contains considerably unequal occurrences of different classes attribute, which is not rare in sentiment analysis (Saad 2014). According to Saad (Saad 2014), Accuracy is a good indicator only if the data is symmetric in the sense that it has almost the same number of false positive and false negative values. Therefore, other metrics should be reported along with the Accuracy. Regarding the second most common metrics (*Recall* and *Precision)*, similar to Accuracy, it is not particularly helpful to use these metrics in isolation. For instance, an algorithm can get perfect Recall if it simply classifies every item into a single class. On the other hand, a precision

score of 1.0 for a class X indicates that all items labelled as belonging to class X do indeed belong to class X. However, it does not show anything about the items that belong to the same class but have not been classified as X (the false positive portion). Having said that, partial representation of the results (e.g. presenting the results using a single metric) is a barrier to replicating studies. It is another challenge in detecting emotions using ML approaches. Thus, researchers and practitioners should consider multiple metrics when reporting their results to provide a comprehensive picture of their model. As a future direction in this regard, we urge researchers to study the available metrics and provide guidelines on the most suitable performance metrics for ML-based emotion detection. In addition, a proposed research direction is to correlate evaluation metrics to the application domain. For instance, Powers M. (Powers 2020) stated that '*Recall tends to be neglected or averaged away in Machine Learning and Computational Linguistics (where the focus is on how confident we can be in the rule or classifier). However, in a Computational Linguistics/Machine Translation context, Recall has been shown to have a major weight in predicting the success of Word Alignment*'. On the other hand, Recall is extremely important in the health domain because the main concern is identifying all actual positive cases. Thus, it is important to consider these differences in future research.

- **Execution and training time is critical; however, it is neglected by many studies**. Our results show that the Time metric is rarely used in the literature. Only four studies reported results about the time (execution time, training time, or testing time). Nonetheless, the Time metric is as important as any other metric. Kukolja et al. (Kukolja et al. 2014), for instance, found that an emotion detection algorithm is highly accurate. However, applying the feature selection method takes a very long time. Therefore, they concluded that this algorithm is good but may not be useful in real life. Hence, it is not recommended during a live interaction with the users. Accordingly, studying the relationship between these metrics (time, accuracy, etc.) is a promising research direction. Another future research direction is to study the relationship between the application domain and evaluation metrics to identify the most important metric for each scenario. The literature needs a study of the tradeoffs between evaluation metrics in various emotion detection domains. For instance, in a domain where active live interactions with users are fundamental, we may prefer algorithm X that classifies faster than algorithm Y, even if Y is slightly more accurate than X. There is some initial research in this direction (Kukolja et al. 2014); however, it is not fully addressed yet, and more research is needed.
- **The impact of data size**. ML algorithms are affected by the data size; normally, the bigger the data available, the better the algorithm's performance (Nugrahaeni and Mutijarsa 2017). However, the literature lacks a study that explores the relationship between the most efficient ML algorithms and data size. Therefore, we invite researchers to invest in this direction and study which algorithm is better for which amount of data – small, medium, and large datasets. For instance, some algorithms are less affected by the data size

5.5. Application area

Our results show that ML algorithms for emotion detection have been applied in more than 15 areas. The most popular areas are social media, Multimedia, and eCommerce. Sentiment analysis is also of special importance for other domains where systems have more extensive interactions with users, such as the learning domain. However, we noticed that the work in these domains is relatively limited and did not get as popularity as the aforementioned domains. Thus, we recommend that more efforts be focused on these domains that have gained lower attention so far, including learning and education, crime, tourism, news, finance and economics, and politics (as presented in Figure 7).

5.6. Future research directions

Despite the large number of research in this domain, several aspects still need further investigation and evaluation. Following are future research directions needed to advance this domain: (1) Section 4.2.2 summarises a comparative analysis of different ML algorithms based on previous work. However, the results in this section need to be further investigated through a **thorough study that provides a comparative analysis of various ML algorithms using benchmarked experimental settings**. (2) our results show that shallow ML algorithms are more common than DL. However, this conclusion cannot be taken in isolation from other factors (as discussed in Section 5.1). Therefore, a **comprehensive study is needed to understand why DL is not as common as ML despite its proven high performance**. (3) our study indicated that English is the most considered language for text-based emotion detection. Therefore, researchers are invited **to explore the effectiveness of different ML algorithms in various**

languages. Also, the literature lacks a study that clearly shows which ML algorithm is more effective for each language and under which settings. (4) Based on the trend in the literature during the last 10 years, we saw increased adoption of DL algorithms and the availability of datasets. We expect that DL will gain more popularity in the coming years. Thus, **a literature review that extends the current one is needed in the near future to explore the changes in the literature**. (5) related to the previous point, it is believed that one of the obstacles in introducing DL approaches is the lack of sufficient data, especially related to modalities other than text (e.g. face recognition, body gestures, etc.). **Providing datasets and making them available for research purposes would advance the research** in this direction and allow more researchers to dig deeper into this domain.

6. Conclusion

This paper provides a systematic literature review of the machine learning approaches used for emotion detection and sentiment analysis. Particularly, it investigates trends, ML techniques and algorithms, available datasets, application domains, and evaluation metrics. The review results show increasing attention toward ML-based emotion detection, and conventional supervised ML approaches (such as SVM and NB) are the most common approaches, and they are among the best-performing algorithms. The review also highlighted challenges, gaps, and future research directions to enhance the development of interactive systems.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grant. The research is conducted as part of the Dalhousie Persuasive Computing Lab.

ORCID

Alaa Alslaity http://orcid.org/0000-0002-1879-9258
Rita Orji http://orcid.org/0000-0001-6152-8034

References

Alkalbani, Asma Musabah, Ahmed Mohamed Ghamry, Farookh Khadeer Hussain, and Omar Khadeer Hussain. 2016. "Predicting the Sentiment of SaaS Online Reviews Using Supervised Machine Learning Techniques." *Proceedings of the International Joint Conference on Neural Networks* 2016-Octob: 1547–1553. doi:10.1109/IJCNN.2016.7727382.

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. 2005. Emotions from Text: Machine Learning for Text-based Emotion Prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 579–586. doi:10.3115/1220575.1220648.

Alslaity, Alaa, and Thomas Tran. 2021. "Users' Responsiveness to Persuasive Techniques in Recommender Systems." *Frontiers in Artificial Intelligence* 4. doi:10.3389/FRAI.2021.679459.

Anjaria, Malhar, and Ram Mohana Reddy Guddeti. 2014. "A Novel Sentiment Analysis of Social Networks Using Supervised Learning." *Social Network Analysis and Mining* 4: 181. doi:10.1007/s13278-014-0181-9.

Appel, Orestes, Francisco Chiclana, Jenny Carter, and Hamido Fujita. 2016. "A Hybrid Approach to the Sentiment Analysis Problem at the Sentence Level." *Knowledge-Based Systems* 108: 110–124. doi:10.1016/j.knosys.2016.05.040.

Ashwin, T. S., Jijo Jose, G. Raghu, and G. Ram Mohana Reddy. 2016. An E-Learning System with Multifacial Emotion Recognition Using Supervised Machine Learning. *Proceedings - IEEE 7th International Conference on Technology for Education, T4E 2015*: 23–26. doi:10.1109/T4E.2015.21.

Bansal, Neetika, and Ashima Singh. 2016. A Review on Opinionated Sentiment Analysis Based Upon Machine Learning Approach. *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016* 2. doi:10.1109/INVENTIVE.2016.7824843.

Becker, Karin, Viviane P. Moreira, and Aline G.L. dos Santos. 2017. "Multilingual Emotion Classification Using Supervised Learning: Comparative Experiments." *Information Processing & Management* 53 (3): 684–704. doi:10.1016/J.IPM.2016.12.008.

Bilgin, Metin, and İzzet Fatih Şentürk. 2017. Sentiment Analysis on Twitter Data with Semi-Supervised Doc2Vec. *2nd International Conference on Computer Science and Engineering, UBMK 2017*: 661–666. doi:10.1109/UBMK.2017.8093492.

Bing, Liu, and Zhang Lei. 2013. *A Survey of Opinion Mining and Sentiment Analysis*. doi:10.1007/978-1-4614-3223-4.

Boiy, Erik, and Moens Marie-Francine. 2009. "A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts." *Information Retrival* 12: 526–558. doi:10.1007/s10791-008-9070-z.

Chai, T., and R. R. Draxler. 2014. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? -Arguments Against Avoiding RMSE in the Literature." *Geoscientific Model Development* 7 (3): 1247–1250. doi:10.5194/GMD-7-1247-2014.

Chen, Zhenpeng, Yanbin Cao, Huihan Yao, Xuan Lu, Xin Peng, Hong Mei, and Xuanzhe Liu. 2021. "Emoji-powered Sentiment and Emotion Detection from Software Developers' Communication Data." *ACM Transactions on Software Engineering and Methodology* 30: 2. doi:10.1145/3424308.

Chen, Mengmeng, Lifen Jiang, Chunmei Ma, and Huazhi Sun. 2019. Bimodal Emotion Recognition Based on Convolutional Neural Network. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC '19*. Accessed May 22, 2021. doi:10.1145/3318299.3318347.

Chiong, Raymond, Marc T.P. Adam, Zongwen Fan, Bernhard Lutz, Zhongyi Hu, and Dirk Neumann. 2018. A Sentiment Analysis-Based Machine Learning Approach for Financial Market Prediction Via News Disclosures. In *GECCO 2018 Companion - Proceedings of the 2018 Genetic and Evolutionary Computation Conference Companion*, 278–279. doi:10.1145/3205651.3205682.

Doma, Vikrant, and Matin Pirouz. 2020. "A Comparative Analysis of Machine Learning Methods for Emotion Recognition Using EEG and Peripheral Physiological Signals." *Journal of Big Data 2020 7:1* 7 (1): 1–21. doi:10.1186/S40537-020-00289-7.

Ekman, Paul. 2008. "An Argument for Basic Emotions." *Cognition and Emotion* 6 (3–4): 169–200. doi:10.1080/02699939208411068.

Elbagir, Shihab, and Jing Yang. 2018. Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*. Accessed May 22, 2021. doi:10.1145/3302425.3302492.

Fontaine, Johnny R.J., Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. 2016. "The World of Emotions is not Two-Dimensional." *Psychological Science* 18 (12): 1050–1057. doi:10.1111/J.1467-9280.2007.02024.X.

Ghosh, Monalisa, and Goutam Sanyal. 2018. "An Ensemble Approach to Stabilize the Features for Multi-Domain Sentiment Analysis Using Supervised Machine Learning." *Journal of Big Data* 5 (1): 1–25. doi:10.1186/s40537-018-0152-5.

Goldberg, Andrew B, and Xiaojin Zhu. 2006. Seeing Stars when there aren't Many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization. In *Proceedings of TextGraphs: The 1st Workshop on Graph-Based Methods for Natural Language Processing*, 45–52. doi:10.5555/1654758.

Goodfellow, Ian, Yoshua Bengio, and Courville Aaron. 2016. *Deep Learning*. London: MIT Press.

Gupta, Rahul, Saurabh Sahu, Carol Espy-Wilson, and Shrikanth Narayanan. 2018. Semi-supervised and Transfer Learning Approaches for Low Resource Sentiment Classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2018-April: 5109–5113. doi:10.1109/ICASSP.2018.8461414.

Hamdan, Hani, Pierre Vigier, and Frédéric Wantiez. 2017. Data Representation in Sentiment Analysis Task. In *In Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 1–6. doi:10.1145/3109761.3158414.

Hammad, Mustafa, and Mouhammd Al-Awadi. 2016. "Sentiment Analysis for Arabic Reviews in Social Networks Using Machine Learning." *Advances in Intelligent Systems and Computing* 448: 131–139. doi:10.1007/978-3-319-32467-8_13.

Han, Zhongmei, Jiyi Wu, Changqin Huang, Qionghao Huang, and Meihua Zhao. 2020. "A Review on Sentiment Discovery and Analysis of Educational Big-Data." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (1): e1328. doi:10.1002/WIDM.1328.

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting Document Knowledge for Aspect-level Sentiment Classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 2: 579–585. Accessed September 9, 2021. https://arxiv.org/abs/1806.04346v1.

Hu, Xia, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised Sentiment Analysis with Emotional Signals. In *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 607–617. doi:10.1145/2488388.2488442.

Imran, Ali Shariq, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. "Cross-cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on Covid-19 Related Tweets." *IEEE Access* 8: 181074–181090. doi:10.1109/ACCESS.2020.3027350.

Indulkar, Yash, and Abhijit Patil. 2021. Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis. *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, July 2014: 295–299. doi:10.1109/ESCI50559.2021.9396925.

Ismail, Rua, Mawada Omer, Mawada Tabir, Noor Mahadi, and Izzeldein Amin. 2018. Sentiment Analysis for Arabic Dialect Using Supervised Learning. *2018 International Conference on Computer, Control, Electrical, and Electronics Engineering, ICCCEEE 2018*: 0–5. doi:10.1109/ICCCEEE.2018.8515862.

Jain, Kruttika, and Shivani Kaushal. 2018. A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis. *2018 7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2018*: 483–487. doi:10.1109/ICRITO.2018.8748793.

Jain, Praphula Kumar, Rajendra Pamula, and Gautam Srivastava. 2021. "A Systematic Literature Review on Machine Learning Applications for Consumer Sentiment Analysis Using Online Reviews." *Computer Science Review* 41. doi:10.1016/J.COSREV.2021.100413.

Jang, Eun Hye, Byoung Jun Park, Sang Hyeob Kim, Myung Ae Chung, Mi Sook Park, and Jin Hun Sohn. 2014. "Emotion Classification Based on bio-Signals Emotion Recognition Using Machine Learning Algorithms." *Proceedings - 2014 International Conference on Information Science, Electronics and Electrical Engineering, ISEEE 2014* 3: 1373–1376. doi:10.1109/InfoSEEE.2014.6946144.

Janiesch, Christian, Patrick Zschech, and Kai Heinrich. 2021. "Machine Learning and Deep Learning." *Electronic Markets* 31 (3): 685–695. doi:10.1007/S12525-021-00475-2/TABLES/2.

Javaid, Mohd, Abid Haleem, Ravi Pratap Singh, Shanay Rab, and Rajiv Suman. 2021. "Internet of Behaviours (IoB) and its Role in Customer Services." *Sensors International* 2: 100122. doi:10.1016/J.SINTL.2021.100122.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science* 349 (6245): 255–260. doi:10.1126/SCIENCE.AAA8415.

Joseph, Lentin, S. Pramod, and Lekha S. Nair. 2018. Emotion Recognition in a Social Robot for Robot-assisted Therapy

to Autistic Treatment using Deep Learning. *Proceedings of 2017 IEEE International Conference on Technological Advancements in Power and Energy: Exploring Energy Solutions for an Intelligent Power Grid, TAP Energy 2017*: 1–6. doi:10.1109/TAPENERGY.2017.8397220.

Kannadaguli, Prashanth, and Vidya Bhat. 2018. "A Comparison of Bayesian and HMM Based Approaches in Machine Learning for Emotion Detection in Native Kannada Speaker." *2018 IEEMA Engineer Infinite Conference, ETechNxT 2018* 1: 1–6. doi:10.1109/ETECHNXT.2018.8385377.

Karim, Mirsa, and Smija Das. 2018. "Sentiment Analysis on Textual Reviews." *IOP Conference Series: Materials Science and Engineering* 396 (1): 122–127. doi:10.1088/1757-899X/396/1/012020.

Kastrati, Zenun, Fisnik Dalipi, Ali Shariq Imran, Krenare Pireva Nuci, and Mudasir Ahmad Wani. 2021. "Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study." *Applied Sciences 2021* 11 (9): 3986. doi:10.3390/APP11093986.

Kaur, Jasleen, and Jatinderkumar R. Saini. 2014. "Emotion Detection and Sentiment Analysis in Text Corpus: A Differential Study with Informal and Formal Writing Styles." *International Journal of Computer Applications* 101 (9): 1–9. doi:10.5120/17712-8078.

Khan, Aurangzeb, Umair Younis, Alam Sher Kundi, Muhammad Zubair Asghar, Irfan Ullah, Nida Aslam, and Imran Ahmed. 2020. Sentiment Classification of User Reviews Using Supervised Learning Techniques with Comparative Opinion Mining Perspective. In *Advances in Intelligent Systems and Computing*, 23–29. doi:10.1007/978-3-030-17798-0_3.

Kitchenham, Barbara. 2004. "Procedures for Performing Systematic Reviews, Version 1.0." *Empirical Software Engineering* 33 (2004): 1–26.

Kukolja, Davor, Siniša Popović, Marko Horvat, Bernard Kovač, and Krešimir Ćosić. 2014. "Comparative Analysis of Emotion Estimation Methods Based on Physiological Measurements for Real-Time Applications." *International Journal of Human-Computer Studies* 72 (10–11): 717–727. doi:10.1016/J.IJHCS.2014.05.006.

Lalata, Jay-ar P, Aurora Blvd, Cubao Quezon City, Bobby Gerardo, and Ruji Medina. 2019. A Sentiment Analysis Model for Faculty Comment Evaluation Using Ensemble Machine Learning Algorithms. In *Proceedings of the 2019 International Conference on Big Data Engineering (BDE 2019) - BDE 2019*. Accessed May 22, 2021. doi:10.1145/3341620.3341638.

Lee, Hy, and C ePulze Sdn Bhd. 2011. "Chinese Sentiment Analysis Using Maximum Entropy." *Proceedings of the Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2011)* 72: 89–93. Accessed May 21, 2021. http://acl.eldoc.ub.rug.nl/mirror/W/W11/W11-37.pdf#page = 105.

Lee, Po Ming, Yun Teng, and Tzu Chien Hsiao. 2012. XCSF for Prediction on Emotion Induced by Image Based on Dimensional Theory Of Emotion. In *GECCO'12 - Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Companion*, 375–382. doi:10.1145/2330784.2330842.

Lin, Chenghua, and Yulan He. 2009. Joint Sentiment/topic Model for Sentiment Analysis. *International Conference on Information and Knowledge Management, Proceedings*: 375–384. doi:10.1145/1645953.1646003.

López, Sergio Altares, and Juan J. Cuadrado-Gallego. 2019. Supervised Learning Methods Application to Sentiment Analysis. In *ACM International Conference Proceeding Series*, 10–12. doi:10.1145/3331076.3331086.

Ly, Son Thai, Guee-Sang Lee, Soo-Hyung Kim, and Hyung-Jeong Yang. 2018. Emotion Recognition via Body Gesture: Deep Learning Model Coupled with Keyframe Selection. In *Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence - MLMI2018*. Accessed May 22, 2021. doi:10.1145/3278312.3278313.

Machado, Samuel, Ana Carolina Ribeiro, and Jorge Oliveira e Sá. 2019. Machine Learning Algorithms and Techniques for Sentiment Analysis in Scientific Paper Reviews: A Systematic Literature Review. In *Atas da Conferencia da Associacao Portuguesa de Sistemas de Informacao*.

Majeed, Adil, Hasan Mujtaba, and Mirza Omer Beg. 2020. Emotion Detection in Roman Urdu Text using Machine Learning. In *Proceedings - 2020 35th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASEW 2020*, 125–130. doi:10.1145/3417113.3423375.

Malheiro, Ricardo, Renato Panda, Paulo J. S. Gomes, and Rui Pedro Paiva. 2013. Music Emotion Recognition from Lyrics: A Comparative Study. *6th International Workshop on Music and Machine Learning – MML 2013 – in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2013*. Accessed September 6, 2021. https://estudogeral.sib.uc.pt/handle/10316/95165.

Mankar, Tejas, Tushar Hotchandani, Manish Madhwani, Akshay Chidrawar, and C. S. Lifna. 2018. Stock Market Prediction based on Social Sentiments using Machine Learning. *2018 International Conference on Smart City and Emerging Technology, ICSCET 2018*. doi:10.1109/ICSCET.2018.8537242.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. "Sentiment Analysis Algorithms and Applications: A Survey." *Ain Shams Engineering Journal* 5 (4): 1093–1113. doi:10.1016/j.asej.2014.04.011.

Mehrabian, Albert. 1996. "Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament." *Current Psychology* 14 (4): 261–292. doi:10.1007/bf02686918.

Mite-Baidal, Karen, Carlota Delgado-Vera, Evelyn Solís-Avilés, Ana Herrera Espinoza, Jenny Ortiz-Zambrano, and Eleanor Varela-Tapia. 2018. "Sentiment Analysis in Education Domain: A Systematic Literature Review." *Communications in Computer and Information Science* 883: 285–297. doi:10.1007/978-3-030-00940-3_21/COVER.

Muhammad, Waqar, Maria Mushtaq, Khurum Nazir Junejo, and Muhammad Yaseen Khan. 2020. "Sentiment Analysis of Product Reviews in the Absence of Labelled Data Using Supervised Learning Approaches." *Malaysian Journal of Computer Science* 33 (2): 118–132. doi:10.22452/mjcs.vol33no2.3.

Muhammad, Ghulam, and M. Shamim Hossain. 2021. "Emotion Recognition for Cognitive Edge Computing Using Deep Learning." *IEEE Internet of Things Journal* 4662 (c), doi:10.1109/JIOT.2021.3058587.

Munezero, Myriam, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. "Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text." *IEEE Transactions on Affective Computing* 5 (2): 101–111. doi:10.1109/TAFFC.2014.2317187.

Naqa, Issam El, and Martin J. Murphy. 2015. "What Is Machine Learning?" *Machine Learning in Radiation Oncology*, 3–11. doi:10.1007/978-3-319-18305-3_1.

Narendra, B., K. Uday Sai, G. Rajesh, K. Hemanth, M. V. Chaitanya Teja, and K. Deva Kumar. 2016. "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies." *International Journal of Intelligent Systems and Applications* 8 (8): 66–70. doi:10.5815/IJISA.2016.08.08.

Neumann, Michael, and Ngoc Thang Vu. 2019. Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*: 7390–7394.

Nugrahaeni, Ratna Astuti, and Kusprasapta Mutijarsa. 2017. Comparative Analysis of Machine Learning KNN, SVM, and Random Forests Algorithm for Facial Expression Classification. *Proceedings - 2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*: 163–168. doi:10.1109/ISEMANTIC.2016.7873831.

Olagunju, Tolulope, Oladapo Oyebode, and Rita Orji. 2020. "Exploring Key Issues Affecting African Mobile ECommerce Applications Using Sentiment and Thematic Analysis." *IEEE Access* 8: 114475–114486. doi:10.1109/ACCESS.2020.3000093.

Oyebode, Oladapo, Felwah Alqahtani, and Rita Orji. 2020. "Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews." *IEEE Access* 8: 111141–111158. doi:10.1109/ACCESS.2020.3002176.

Pitogo, Vicente A, Christine Diane, and L. Ramos. 2021. Social Media Enabled e-Participation: A Lexicon-based Sentiment Analysis using Unsupervised Machine Learning CCS CONCEPTS. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, 23–25. Accessed May 22, 2021. doi:10.1145/3428502.3428581.

Plutchik, Robert. 1980. A General Psychoevolutionary Theory of Emotion. In *Theories of Emotion*. Academic Press, 3–33. doi:10.1016/b978-0-12-558701-3.50007-7.

Pong-Inwong, Chakrit, and Konpusit Kaewmak. 2017. Improved Sentiment Analysis for Teaching Evaluation Using Feature Selection and Voting Ensemble Learning Integration. *2016 2nd IEEE International Conference on Computer and Communications, ICCC 2016 - Proceedings*: 1222–1225. doi:10.1109/CompComm.2016.7924899.

Poornima, A., and K. Sathiya Priya. 2020. A Comparative Sentiment Analysis of Sentence Embedding Using Machine Learning Techniques. *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*: 493–496. doi:10.1109/ICACCS48705.2020.9074312.

Powers, David M. W. 2020. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Accessed September 6, 2021. https://arxiv.org/abs/2010.16061v1.

Qian, Jun, Zhendong Niu, and Chongyang Shi. 2018. "Sentiment Analysis Model on Weather Related Tweets with Deep Neural Network." *In ACM International Conference Proceeding Series*, 31–35. doi:10.1145/3195106.3195111.

Rahman, Shaomi, Jonayed Nafis Hemel, Syed Junayed Ahmed Anta, Hossain Al Muhee, and Jia Uddin. 2019. Sentiment Analysis using R: An Approach to Correlate Cryptocurrency Price Fluctuations with Change in User Sentiment using Machine Learning. *2018 Joint 7th International Conference on Informatics, Electronics and Vision and 2nd International Conference on Imaging, Vision and Pattern Recognition, ICIEV-IVPR 2018* 2021, January 2021: 492–497. doi:10.1109/ICIEV.2018.8641075.

Roh, Yuji, Geon Heo, and Steven Euijong Whang. 2021. "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective." *IEEE Transactions on Knowledge and Data Engineering* 33 (4): 1328–1347. doi:10.1109/TKDE.2019.2946162.

Rohini, V., Merin Thomas, and C. A. Latha. 2017. Domain Based Sentiment Analysis in Regional Language-Kannada using Machine Learning Algorithm. *2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings*: 503–507. doi:10.1109/RTEICT.2016.7807872.

Rudin, Cynthia. 2019. "Stop Explaining Black box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead." *Nature Machine Intelligence 2019 1:5* 1 (5): 206–215. doi:10.1038/s42256-019-0048-x.

Russell, James A. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39 (6): 1161–1178. doi:10.1037/H0077714.

Russell, Stuart, and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach.* Pearson. https://www.pearson.com/us/higher-education/program/Russell-Artificial-Intelligence-A-Modern-Approach-4th-Edition/PGM1263338.html.

Saad, Farag. 2014. "Baseline Evaluation: An Empirical Study of the Performance of Machine Learning Algorithms in Short Snippet Sentiment Analysis." *In ACM International Conference Proceeding Series*, doi:10.1145/2637748.2638420.

Sagnika, Santwana, Anshuman Pattanaik, Bhabani Shankar Prasad Mishra, and Saroj K. Meher. 2020. "A Review on Multi-Lingual Sentiment Analysis by Machine Learning Methods." *Journal of Engineering Science and Technology Review* 13 (2): 154–166. doi:10.25103/JESTR.132.19.

Sajib, Mahamudul Islam, Shoeib Mahmud Shargo, and Md Alomgir Hossain. 2019. Comparison of the Efficiency of Machine Learning Algorithms on Twitter Sentiment Analysis of Pathao. *2019 22nd International Conference on Computer and Information Technology, ICCIT 2019*: 18–20. doi:10.1109/ICCIT48885.2019.9038208.

Sarkar, Pritam, and Ali Etemad. 2020. "Self-supervised ECG Representation Learning for Emotion Recognition." *IEEE Transactions on Affective Computing* 3045 (c): 1–13. doi:10.1109/TAFFC.2020.3014842.

Shah, Parita, and Priya Swaminarayan. 2022. "Machine Learning-Based Sentiment Analysis of Gujarati Reviews." *International Journal of Data Analysis Techniques and Strategies* 14 (2): 105–121. doi:10.1504/IJDATS.2022.10049552.

Sharma, Anuj, and Shubhamoy Dey. 2012. A Comparative Study of Selection and Machine Learning Techniques for Sentiment Analysis. In *Proceeding of the 2012 ACM Research in Applied Computation Symposium, RACS 2012*, 1–7. doi:10.1145/2401603.2401605.

Singh, V. K., R. Piryani, A. Uddin, and P. Waila. 2013. Sentiment Analysis of Movie Reviews: A New feature-based Heuristic for Aspect-level Sentiment Classification. *Proceedings - 2013 IEEE International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing, iMac4s 2013*: 712–717. doi:10.1109/IMAC4S.2013.6526500.

Singh, Sudheer Kumar, Prabhat Verma, and Pankaj Kumar. 2020. "Sentiment Analysis Using Machine Learning Techniques on Twitter: A Critical Review." *Advances in Mathematics: Scientific Journal* 9 (9): 7085–7092. doi:10.37418/AMSJ.9.9.58.

Song, Tongshuai, Guanming Lu, and Jingjie Yan. 2020. Emotion Recognition Based on Physiological Signals Using Convolution Neural Networks. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*. Accessed May 22, 2021. doi:10.1145/3383972.3384003.

Soumya, S., and K. V. Pramod. 2020. "Sentiment Analysis of Malayalam Tweets Using Machine Learning Techniques." *ICT Express* 6 (4): 300–305. doi:10.1016/j.icte.2020.04.003.

Studiawan, Hudan, Ferdous Sohel, and Christian Payne. 2020. "Anomaly Detection in Operating System Logs with Deep Learning-Based Sentiment Analysis." *IEEE Transactions on Dependable and Secure Computing* XX (XX): 1–13. doi:10.1109/TDSC.2020.3037903.

Sultana, Jabeen, Nasreen Sultana, Kusum Yadav, and Fayez Alfayez. 2018. Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach. *21st Saudi Computer Society National Computer Conference, NCC 2018*: 14–18. doi:10.1109/NCG.2018.8593108.

Tang, Huifeng, Songbo Tan, and Xueqi Cheng. 2009. "A Survey on Sentiment Detection of Reviews." *Expert Systems with Applications* 36 (7): 10760–10773. doi:10.1016/j.eswa.2009.02.063.

Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. 2016. "Classification of Sentiment Reviews Using n-Gram Machine Learning Approach." *Expert Systems with Applications* 57: 117–126. doi:10.1016/j.eswa.2016.03.028.

Vaish, Abhishek, and Pinki Kumari. 2014. "A Comparative Study on Machine Learning Algorithms in Emotion State Recognition Using ECG." *Advances in Intelligent Systems and Computing* 236: 1467–1476. doi:10.1007/978-81-322-1602-5_147.

Van Huynh, Thong, Hyung Jeong Yang, Guee Sang Lee, Soo Hyung Kim, and In Seop Na. 2019. Emotion Recognition by Integrating Eye Movement Analysis and Facial Expression Model. In *ACM International Conference Proceeding Series*, 166–169. doi:10.1145/3310986.3311001.

Vaseeharan, Thinesharan, and Achala Aponso. 2020. Review on Sentiment Analysis of Twitter Posts about News Headlines Using Machine Learning Approaches and Naïve Bayes Classifier. In *12th International Conference on Computer and Automation Engineering (ICCAE 2020)*, 33–37. doi:10.1145/3384613.3384650.

Vijayakumar, Sowmya, Ronan Flynn, and Niall Murray. 2020. A Comparative Study of Machine Learning Techniques for Emotion Recognition from Peripheral Physiological Signals. *2020 31st Irish Signals and Systems Conference, ISSC 2020*. doi:10.1109/ISSC49989.2020.9180193.

Watson, David, Lee Anna Clark, and Auke Tellegen. 1988. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54 (6): 1063–1070.

Weichselbraun, Albert, Stefan Gindl, and Arno Scharl. 2010. "A Context-Dependent Supervised Learning Approach to Sentiment Detection in Large Textual Databases." *Journal of Information and Data Management* 1 (3): 329–342. Accessed May 21, 2021. www.tripadvisor.com.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 347–354. doi:10.3115/1220575.1220619.

Xu, Xinzhou, Jun Deng, Eduardo Coutinho, Chen Wu, Li Zhao, and Björn W. Schuller. 2019. "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition." *IEEE Transactions on Multimedia* 21 (3): 795–808. doi:10.1109/TMM.2018.2865834.

Yiran, Ye, and Sangeet Srivastava. 2019. Aspect-based Sentiment Analysis on Mobile Phone Reviews with LDA. In *ACM International Conference Proceeding Series*, 101–105. doi:10.1145/3340997.3341012.

Yu, Tian, Izak Benbasat, and Ronald T. Cenfetelli. 2011. Toward deep Understanding of Persuasive Product Recommendation Agents. In *International Conference on Information Systems 2011, ICIS 2011*, 1832–1840.

Zhang, Ying, and Chen Ling. 2018. "A Strategy to Apply Machine Learning to Small Datasets in Materials Science." *npj Computational Materials 2018 4:1* 4 (1): 1–8. doi:10.1038/s41524-018-0081-z.

Zhang, Cong, Hui Lu, and Zekun Song. 2017. Sentiment Analysis for Chinese Version based on Machine Learning. In *Proceedings of the 6th International Conference on Information Engineering - ICIE '17*. Accessed May 22, 2021. doi:10.1145/3078564.3078573.

Zhang, Xueying, and Xianghan Zheng. 2017. Comparison of Text Sentiment Analysis based on Machine Learning. *Proceedings - 15th International Symposium on Parallel and Distributed Computing, ISPDC 2016*: 230–233. doi:10.1109/ISPDC.2016.39.

Zhuang, Chen, and Qian Tieyun. 2019. "Transfer Capsule Network for Aspect Level Sentiment Classification." *Annual Meeting of the Association for Computational Linguistics* 57: 547–556.

Appendix A. Comprehensive overview of the research in machine learning approaches for sentiment analysis and emotion detection

Ref	Year	Application Area	ML technique	Algorithm	Modality	Text Dimension	Language	Datasets	Evaluation Metrics
(Ashwin et al. 2016)	2016	Social Media & Blogs, Multimedia	Supervised	NB, Maximum Entropy	Text	Sentence	English	Sentiment140, Movie review	Accuracy, Precision, Recall, F1
(Indulkar and Patil 2021)	2009	Social Media & Blogs	Supervised	, Multimial NB, Maximum Entropy, SVM	Text	Sentence	English, Dutch, French	Self-gathered	Accuracy, Precision, Recall
(Appel et al. 2016)	2014	Social Media & Blogs	Supervised	NB, Maximum Entropy, SVM, ANN	Text	Sentence	English	Self-gathered	Accuracy
(Rohini, Thomas, and Latha 2017)	2019	Learning	Supervised	NB, Logistic Regression, SVM, Decision Tree, Random Forest, Others	Text	Sentence	English	Self-gathered	Accuracy, Recall, F1
(Ghosh and Sanyal 2018)	2018	Finance & Economic	Supervised	SVM	Text	Document	English	German ad hoc announcements in English	Accuracy, Time
(Jain and Kaushal 2018)	2018	eCommerce	Supervised	Bernoulli NB, Multimial NB, Logistic Regression, SVM, Random Forest	Text	Feature	English	Self-gathered	, Precision, Recall, F1
(Goodfellow, Bengio, and Aaron 2016)	2019	General/Multiple	Supervised	Gaussian processes reg ANN	Speech	NA		International Affective Digitized Sound (IADS)	, RMSE, R square
(Boiy and Marie-Francine 2009)	2019	Multimedia	Deep Learning	CNN	Multi modals	NA		eNTERFACE'05	Accuracy
(Rudin 2019)	2011	eCommerce	Supervised	Maximum Entropy	Text	Sentence	Chinese	Self-gathered	Accuracy
[149]	2016	Multimedia	Supervised	NB, Maximum Entropy, SVM, SGD	Text	Sentence	English	IMDB	Accuracy, Precision, Recall, F1
(Hu et al. 2013)	2018	Social Media & Blogs	Supervised	NB, SVM	Text	Document	Arabic	Multiple previously collected tweets DS	Accuracy
(Karim and Das 2018)	2017	Multimedia	Supervised	Logistic Regression, Random Forest	Text	Document	English	IMDB	Accuracy
(Jain, Pamula, and Srivastava 2021)	2011	eCommerce	Deep Learning	DNN	Text	Document	English	Amazon reviews	, Transfer loss
(Sharma and Dey 2012)	2020	General/Multiple	Supervised	SVM, Decision Tree, Random Forest, KNN	Text	Sentence	Romanian, Urdo	Self-gathered	Accuracy, Precision, Recall, F1
[140]	2020	General/Multiple	Deep Learning	CNN	Physiological signals	NA	NA	MAHNOB-HCI	Accuracy
(Lee and Sdn Bhd 2011)	2019	Multimedia	Supervised	ANN	Body gestures	NA	NA	Acted Facial Expressions in the Wild (AFEW)	Accuracy
(Sarkar and Etemad 2020)	2018	Multimedia	Deep Learning	CNN	Body gestures	NA	NA	FABO	Accuracy
(Alslaity and Tran 2021)	2005	Learning	Supervised	NB	Text		English	Self-gathered	Accuracy, Precision, Recall, F1
[154]	2012	Multimedia	Supervised	NB, SVM	Text	Document	English	Polarity DS,, Self-gathered	Accuracy, Precision, Recall, F1
(Kastrati et al. 2021)	2020	,Multimedia	Supervised	NB, Logistic Regression, SVM, Random Forest	Text	Document	English	Self-gathered	, Precision, Recall, F1
(Gupta et al. 2018)	2016	Social Media & Blogs	Supervised	NB, Maximum Entropy, SVM	Text	Document	English	Niek j. Sanders, Polarity	Accuracy

(Sagnika et al. 2020)	2009	Multimedia	Unsupervised	Others	Text	Document	English	Movie Review DS	Accuracy
(Khan et al. 2020)	2008	Multimedia	Supervised	NB, Maximum Entropy, SVM	Speech	NA	Chinese	YY Music Group	Accuracy
[145]	2014	Social Media & Blogs	Supervised	SVM, Others	Text	Document	English	Self-gathered	, F1
(Pitogo et al.)	2019	General/Multiple	semi-supervised	Others	Text	Document	English	Cornell movie review, Amazon product review	Accuracy
(Rahman et al. 2019)	2018	General/Multiple	Deep Learning	CNN	Body gestures	NA	NA	The Extended Cohn-Kanade Dataset (CK+)	Accuracy
[133]	2019	eCommerce	Supervised	, Multimial NB, SVM, CNN, RNN, Random Forest, Others	Text	Document	English	Self-gathered	Accuracy, Precision, Recall, F1
(Chai and Draxler 2014)	2017	News	Supervised	NB, SVM, CNN	Text	Document	English, Portuguese	SemEvalNews, BRNews	, F1
(Pong-Inwong and Kaewmak 2017)	2017	, Multimedia	semi-supervised	CNN	Multi modals	NA	NA	"AFEW 6.0,	Accuracy
(Vaseeharan and Aponso 2020)	2013	General/Multiple	Supervised	SVM	Text	Self-gathered			, F1
(Vaish and Kumari 2014)	2015	General/Multiple	Supervised	NB, SVM, Random Forest, SGD	Text	Document			Accuracy
(Bansal and Singh 2016)	2019	General/Multiple	Supervised	NB, SVM, Decision Tree, KNN	Text	Feature	English	Self-gathered	, Precision, Recall, F1
(Han et al. 2020)	2007	Social Media & Blogs	Supervised	NB, SVM	Text	Document	English	rotten tomatoes database	Accuracy
(Alm, Roth, and Sproat 2005)	2016	Technology	Supervised	NB, SVM, Decision Tree, KNN	Text	Sentence	English	ISEAR	Accuracy, Precision, Recall
(Elbagir and Yang 2018)	2021	General/Multiple	Supervised	Ensemble	Body gestures	Sentence	English	Self-gathered	Accuracy
[159]	2015	Social Media & Blogs	Deep Learning	Ensemble	Text	Sentence	English	Self-gathered	, Precision, Recall
(Bilgin and Şentürk 2017)	2019	eCommerce	Supervised	SVM	Text	NA	NA	CMU Multi-PIE Face	Accuracy, Precision, Recall
(Janiesch, Zschech, and Heinrich 2021)	2006	, Multimedia	semi-supervised	Others	Text	Sentence	Chinese	Self-gathered	Accuracy
[131]	2020	General/Multiple	Deep Learning	CNN	Biosignals	Sentence	Urdu	publicly available	Accuracy, F1
(Kannadaguli and Bhat 2018)	2018	eCommerce, Multimedia	semi-supervised	Others	Text	Sentence	English	moview review	Accuracy
(Naqa and Murphy 2015)	2012	Tourism	Supervised	NB	Text	NA	NA	AMIGOS, DREAMER, WESAD, SWELL	Accuracy, Precision, Recall
(Ly et al. 2018)	2018	Social Media & Blogs	Supervised	NB, Logistic Regression, SVM, KNN	Text	Sentence	English	UCI sentiment labelled sentences, Movie review dataset	Accuracy, F1
[164]	2017	Social Media & Blogs	Deep Learning	, CNN	Text	Document	English	Self-gathered	Accuracy, Recall
[117]	2018	Social Media & Blogs	Deep Learning	, CNN, DNN	Text	Sentence	Arabic	Self-gathered	, Precision, Recall, F1
(Imran et al. 2020)	2019	eCommerce	Supervised	NB	Text	Document	Chinese	Self-gathered	Accuracy
(Watson, Clark, and Tellegen 1988)	2020	eCommerce	Supervised	NB, Logistic Regression, SVM, Random Forest, KNN	Text	Sentence	English	Self-gathered	Accuracy, Precision, Recall
(Van Huynh et al. 2019)	2018	Social Media & Blogs	Supervised	Bernoulli NB, Multimial NB, SVC, SGD	Text	Sentence	Bengali	Self-gathered	Accuracy, Precision, Recall, F1
(Chen et al. 2021)	2017	Social Media & Blogs	semi-supervised	, Linear Regression, Others	Text	Document	English	Self-gathered	Accuracy
[115]	2009	General/Multiple	Multiple	, Ensemble	Text	Document	English	NLTK	, F1

(*Continued*)

Appendix A. Continued.

Ref	Year	Application Area	ML technique	Algorithm	Modality	Text Dimension	Language	Datasets	Evaluation Metrics
(Oyebode, Alqahtani, and Orji 2020)	2020	Technology	Supervised	NB, Logistic Regression, SVM, Decision Tree, Random Forest, KNN	Text	Document	English, Turkish	Self-gathered	Accuracy, Precision, Recall, F1
[116]	2018	General/Multiple	Supervised	, Logistic Regression, SVM, Decision Tree, KNN	Body gestures	Document	English	Self-gathered	Accuracy, Precision, Recall, F1
[110]	2020	Social Media & Blogs	unsupervised	, Others	Text	Sentence	English	Self-gathered	Accuracy
(Saad 2014)	2018	General/Multiple	Deep Learning	, CNN	Speech	NA	NA	AffectNet	Accuracy
[109]	2020	General/Multiple	Deep Learning	, CNN	Multi modals	Sentence	English	Self-gathered	Accuracy, Others
(Sajib, Shargo, and Hossain 2019)	2019	Social Media & Blogs	Supervised	NB, Logistic Regression, SVM, Decision Tree, Random Forest, Others	Text	Sentence	English	IEMOCAP	, Precision, Recall, F1, False Positive Rate, AUC
[108]	2002	, Multimedia	Supervised	NB, Maximum Entropy, SVM	Text	NA	NA	emotion dataser, RAVDESS dataser	Accuracy
[138]	2014	Social Media & Blogs	Supervised	Ensemble	Text	Sentence	English	yelp labelled, amazon cells labelled	Precision, Recall, F1
(Roh, Heo, and Whang 2021)	2015	Social Media & Blogs	Supervised	Ensemble	Text	Document	English	IMDB	Accuracy, Precision, Recall, F1
[136]	2016	Social Media & Blogs	Supervised	NB, SVM	Text	Sentence	English	Multiple previously collected tweets DS	Accuracy
(Lalata et al. 2019)	2013	Social Media & Blogs	Unsupervised	, Others	Text	Sentence	English	SemEval	Accuracy
(Narendra et al. 2016)	2018	General/Multiple	Supervised	, Others	Speech	Sentence	English	Stanford Dataset	Emotion Error Rate
(Shah and Swaminarayan 2022)	2018	General/Multiple	Deep Learning	, CNN	Body gestures	Sentence	English	Stanford Dataset, Obama-McCain	Others
(Jang et al. 2014)	2020	Social Media & Blogs	Deep Learning	, Logistic Regression, SVM, RNN, Random Forest	Text	Document	Kannada	Self-gathered	Accuracy
[123]	2020	General/Multiple	Supervised	, Gaussian Naive Ba, Others, Random Forest	Text	NA	NA	Fer2013	Accuracy
(Doma and Pirouz 2020)	2019	Health	Supervised	, SVM	Multi modals	Sentence	Hindi	Self-gathered	Accuracy
(Malheiro et al. 2013)	2014	General/Multiple	Supervised	NB, SVM, Others	Biosignals	Document	English	Self-gathered	Accuracy
(Becker, Moreira, and dos Santos 2017)	2015	Learning	Supervised	, SVM	Body gestures	NA	NA	Self-gathered	Accuracy
[153]	2019	Politics	Supervised	NB, SVM	Text	NA	NA	Self-gathered	Accuracy
[142]	2020	Technology	Deep Learning	, RNN	Text	NA	NA	LFW, FDDB, YFD	Accuracy, Precision, Recall, F1
[147]	2018	Health	Deep Learning	, DNN, KNN	Speech	Document	English	Self-gathered	
(Plutchik 1980)	2020	, Social Media & Blogs	Deep Learning	, RNN	Text	Sentence	English	Multiple previously collected OS logs	Accuracy, Precision, Recall
(Munezero et al. 2014)	2017	Politics	Supervised	NBBernoulli NB, Multimial NB, Linear Regression, SVC, SGD	Text	Document	English	Self-gathered	Accuracy, Others
[137]	2016	Social Media & Blogs	Supervised	, Ensemble	Text	Document	English	Twitter US Airline Sentiment dataset	Accuracy
(López and Cuadrado-Gallego 2019)	2021	Social Media & Blogs	Supervised	, Multimial NB, Logistic Regression, Random Forest	Text	Document	English	Self-gathered	Accuracy, Others
[165]	2017	General/Multiple	Supervised	, SVM, Others	Text	Sentence	English	Stanford sentiment140	Accuracy, Time
[130]	2019	Social Media & Blogs	Supervised	NB, Logistic Regression, SVM	Text	Sentence	English	Uber, Ola datasets	Accuracy

[158]	2019	General/Multiple	unsupervised	, Others	Speech	Sentence	Chinese	Self-gathered	Accuracy
[163]	2019	General/Multiple	Supervised	, Others	Speech	Sentence	English	Self-gathered	Accuracy
(Bing and Lei 2013)	2017	General/Multiple	Deep Learning	, DNN	Body gestures	Document	English	GEMEP, ABC, VAM, and eNTERFACE	, Others
(Weichselbraun, Gindl, and Scharl 2010)	2019	, Social Media & Blogs	Deep Learning	, DNN	Text	Document	German, Mandarin	FAU Aibo Emotion German ds, CNDB Mandarin DS	Accuracy, F1
(Studiawan, Sohel, and Payne 2020)	2021		Deep Learning	NB, SVM, CNN, DNN	Text	NA	NA	RaFD, CK+, JAFFE	Accuracy
[144]	2019	, Social Media & Blogs	Supervised	, Random Forest	Text	Document	Hindi, English	Self-gathered	Accuracy, Precision, Recall, F1, Emotion Error Rate, Others
[121]	2021	eCommerce	Supervised	, SVM, Random Forest, KNN	Text	Document	Arabic	Self-gathered	Accuracy
[132]	2019	General/Multiple	Supervised	, Logistic Regression	Text	Document, Sentence	Bangla	Self-gathered	, F1, Time
[125]	2017	, Multimedia	Supervised	, Decision Tree	Text	Document	English	Amazon reviews	, Precision, Recall
(Mehrabian 1996)	2019	General/Multiple	Deep Learning	, RNN	Biosignals	Sentence	Kazakh	Self-gathered	Accuracy
(Majeed, Mujtaba, and Beg 2020)	2014	General/Multiple	Supervised	NB, SVM, Others	Biosignals	Document	Kannada	Self-gathered	Accuracy
(Vijayakumar, Flynn, and Murray 2020)	2020	Technology	Deep Learning	, CNN	Body gestures	NA	NA	Self-gathered	Accuracy, Others
(Soumya and Pramod 2020)	2020	General/Multiple	Deep Learning	, DNN	Speech	NA	NA	Self-gathered	Accuracy
(Muhammad et al. 2020)	2017	Technology	Deep Learning	, CNN	Body gestures	NA	NA	JAFFE, CK+	Accuracy, Time
(Russell and Norvig 2020)	2018	Health	Deep Learning	, Others	Biosignals	NA	Romanian	SRoL	Accuracy, F1, MSE
(Machado, Ribeiro, and e Sá 2019)	2020	General/Multiple	Deep Learning	, CNN	Body gestures	NA	NA	JAFFE, CK+	Accuracy
[112]	2016	Learning	unsupervised	, Ensemble	Text	NA	NA	Self-gathered	Accuracy
(Yu, Benbasat, and Cenfetelli 2011)	2019	General/Multiple	Deep Learning	, CNN	Speech	NA	NA	"FERC-2013	Accuracy, Recall, Others
(Song, Lu, and Yan 2020)	2020	Crime	Supervised	, Multimial NB, SVM, Random Forest, KNN	Text	JAFFE"			Accuracy
(Jordan and Mitchell 2015)	2019	General/Multiple	Supervised	, Others	Text	Sentence	English	Self-gathered	Accuracy, Recognition loss
(Tripathy, Agrawal, and Rath 2016)	2020	General/Multiple	Deep Learning	, CNN, Decision Tree, Random Forest, Others	Speech	Document	English	IEMOCAP, MSP-IMPROV	Accuracy
[143]	2018	Learning	Deep Learning	NB, Logistic Regression, SVM, Decision Tree, Random Forest, Others	Text	Sentence	English	Self-gathered	Accuracy, Precision, Recall, AUC
(Hamdan, Vigier, and Wantiez 2017)	2019	eCommerce	Supervised	, Logistic Regression, SVM, KNN	Text	Sentence	English	Yelp, IMDB	Accuracy
(Mite-Baidal et al. 2018)	2019	, Multimedia	Supervised	, Logistic Regression, SVM, Decision Tree, Random Forest, KNN	Text	Sentence	Thai	SAVEE, RAVDESS, TESS, CREMA-D	Accuracy, Precision, Recall, F1
(Goldberg and Zhu 2006)	2020	General/Multiple	semi-supervised	CNN	Body gestures	Sentence	English	Kalboard 360	RMSE, Others
[167]	2015	, Multimedia	Deep Learning	RNN, Decision Tree	Text	Document	English	Amazon reviews	Accuracy, Precision, Recall, F1
(Zhang and Zheng 2017)	2018	News	Deep Learning	, ANN	Text	Document	English	IMDB	Accuracy, Others
[141]	2020	, Social Media & Blogs	Supervised	NB, SVM, Random Forest	Text	NA	NA	MAHNOB-HCI, INHA	Accuracy, Precision, Recall, F1

(*Continued*)

Appendix A. Continued.

Ref	Year	Application Area	ML technique	Algorithm	Modality	Text Dimension	Language	Datasets	Evaluation Metrics
(Xu et al. 2019)	2018	, Social Media & Blogs	Supervised	SVM, Random Forest	Text	Document	English	IMDB	Accuracy, Precision, Recall, F1
(Alkalbani et al. 2016)	2019	, Social Media & Blogs	Supervised	NB, ANN, CNN, RNN, Decision Tree, Random Forest, Ensemble	Text	Sentence	English	TVKU	Accuracy, Precision, Recall
(Kaur and Saini 2014)	2017	Health	Supervised	NB, Logistic Regression, SVM	Text	Document	Malayalam	Self-gathered	Accuracy
(Nugrahaeni and Mutijarsa 2017)	2020	, Social Media & Blogs	Supervised	NB, Logistic Regression, Maximum Entropy, SVM, Decision Tree, Random Forest, SGD	Text	Sentence	hindi	Self-gathered	Precision, Recall, F1
(Joseph, Pramod, and Nair 2018)	2019	, Social Media & Blogs	Supervised	, Logistic Regression, SVM, ANN, Decision Tree	Text	Document	English	Self-gathered	Accuracy
(Kukolja et al. 2014)	2019	Tourism	Supervised	NBBernoulli NB, Multimial NB, Logistic Regression, SVC	Text	Sentence	English	Self-gathered	Accuracy
[118]	2019	Multimedia	Supervised	Bernoulli NB, Multimial NB, Logistic Regression, SVM, Decision Tree	Text	Sentence	English	Twitter dataset	Accuracy, Precision, Recall, F1
(Neumann and Vu 2019)	2018	Multimedia	Supervised	NB, SVM, Others	Text	Document	English	Sentiment140	Accuracy, Precision, Recall, F1
[150]	2020	Social Media & Blogs	Unsupervised	CNN	Text	Document	English	Yelp	Accuracy
(Powers 2020)	2018	Social Media & Blogs	Deep Learning	CNN	Text	Document	English	Self-gathered	Accuracy, Others
[134]	2019	Social Media & Blogs	Deep Learning	NB, SVM, Random Forest	Text	Document	English	"Cornell, Self-gathered	Accuracy, Precision
[148]	2019	Multimedia	Supervised	CNN	Text	Document	NA	NA	Accuracy
[122]	2019	General/Multiple	Deep Learning	SVM, KNN	Speech	Document	English	Sentiment140	Accuracy
[120]	2017	Finance & Economic	Supervised	NB, SVM	Text	Document	English	tweets	Accuracy, Precision, Recall,
(Lee, Teng, and Hsiao 2012)	2020	General/Multiple	Supervised	DNN	Bio signals	Document	English	Self-gathered	Accuracy
[162]	2020	Social Media & Blogs	Deep Learning	NB, SVM, RNN, Random Forest	Text	Document	English	IMDB	Accuracy