IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 13, Issue 8, August 2025

DOI: 10.17148/IJIREEICE.2025.13831

An inclusive survey on Sentiment analysis: Approaches, Challenges and Trends

Amandeep Kaur

Assistant Professor, Department of Computer Engineering and Technology,
Guru Nanak Dev University, Amritsar, Punjab, India

Abstract: Sentiment analysis, often known as opinion mining, is the process of determining and analyzing people's feelings and viewpoints toward goods, subjects, services, or occasions. Organizations may improve their services and products and create a feedback loop that continuously enhances user experiences by turning such information into relevant knowledge. For sentiment analysis applications, social media platforms and e-commerce websites are crucial since they are significant sources of data that are rich in opinions. Around the world, businesses, governments, and scholars use sentiment analysis to gather commercial insights, assess how the public views policies, and aid in well-informed decision-making. In order to introduce readers to this potent technology and promote additional contributions to the area, this paper provides a thorough description of the tasks, current trends, methodology, and challenges of sentiment analysis.

Keywords: Sentiment Analysis, Opinion Mining, Natural Language Processing (NLP), Machine Learning, Deep Learning, Social Media Analysis, Business Intelligence.

I. INTRODUCTION

Sentiment analysis is the process of extracting sentiments and opinions from textual data, which can exist in various forms, and is considered an important task in Natural Language Processing (NLP)[1]. Beyond textual data, emerging techniques have begun to incorporate multimodal information such as visual data, expanding the scope of sentiment understanding. At its core, sentiment analysis deals with studying people’s opinions, emotions, attitudes, and sentiments toward products, services, issues, events, and topics. Technically, it often becomes a text classification problem, where computational algorithms determine whether a given text conveys a positive, negative, or neutral sentiment.

While sentiment analysis may appear straightforward, in practice, it is an arduous and complex task. This complexity arises due to the necessity of handling multiple NLP subtasks, including sarcasm detection, subjectivity analysis, and context interpretation. Furthermore, real-world text rarely appears in an organized or grammatically correct manner (as in books or newspapers). Instead, data from blogs, forums, e-commerce reviews, and social networking sites often contain orthographic errors, idiomatic expressions, slang, abbreviations, and informal structures, making it much harder to process and derive accurate conclusions.

The importance of sentiment analysis has grown exponentially with the rapid expansion of the internet and social media platforms. Increased internet usage by individuals, organizations, and governments has resulted in massive volumes of opinion-rich data, which must be processed to extract meaningful insights. As people worldwide express their views through forums, social networks, blogs, and product reviews, sentiment analysis has emerged as a critical tool for monitoring public opinion, business intelligence, policy evaluation, and decision-making[2].

This paper provides a comprehensive study of sentiment analysis from multiple perspectives, covering different types of analysis, application domains, methodological processes, and associated challenges. The aim is to familiarize readers with this expanding field, equip them with foundational knowledge, and encourage contributions toward advancing sentiment analysis research.

II. LEVELS OF SENTIMENT ANALYSIS

The sentiments and opinions that are drawn from sentiment analysis are mainly detected at 3 different levels. Namely, document level, sentence level, and aspect level.[3,4,5] This is clearly shown in Figure 1

DOI: 10.17148/IJIREEICE.2025.13831

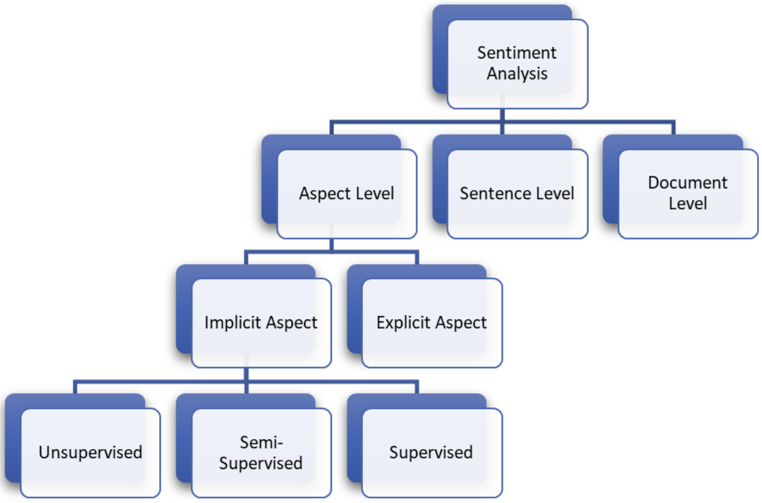

fig1:Levels of Sentiment Analysis

A. Aspect-level sentiment analysis

This level of sentiment analysis is also known as feature-level analysis, as this is a fine-grained model of sentiment analysis. This model represents the opinion of a specific aspect of a given product, service, or entity. For conducting this type of analysis, it is required to extract the entities and their equivalent aspects/features from the given opinionated reviews. This is used to determine the polarity of opinion, which is then summarized and visualized. This level of analysis is currently used by many companies to identify the current requirements of their customers, what specific components or aspects of the products are in demand, and how they can improve their products for increased profit and the sustainability of their company.[6]

For example, the camera of the Google Pixel smartphone is outstanding, which mostly focuses on the software post-processing for the image quality. Reviewers and the general public wrote good reviews about the phone’s camera. Here review is on camera, which is a feature of the Pixel phone. So, this type of analysis shows that people liked the camera, and the analysis showed the review as positive. This approach focuses on the entity rather than a paragraph or a document. It helps to understand a certain feature of a product or service rather than the whole of it.

B. Sentence-level Analysis

This level focuses on the sentence rather than any aspect of the product. The primary objective of this is to predict the opinion of a sentence, whether it is positive, negative, or neutral[7]. This level of analysis is closely related to subjectivity classification, which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions. However, subjectivity is not equivalent to sentiment, as many objective sentiments can imply opinions. Both document-level and sentence-level sentiment analysis are important and useful, but cannot state the true and detailed opinion of all the aspects of an entity, and cannot state what people actually want or feel.

C. Document-level sentiment analysis

This level parses the whole of the document to extract the sentiment of the document, may it is a book, article, blog, or news article, whether it is expressing a positive or negative, or neutral opinion, considering it as a single entity. It is best to choose a document that has been written by a single author or a writer to avoid the mix-up of their thoughts or opinions that evaluate or compare multiple entities. [8]

III. SENTIMENT ANALYSIS PRE-PROCESSING

Sentiment analysis is a big suitcase of natural language processing (NLP) problems. Sentiment analysis has long been mistaken for the task of polarity detection. This, however, is just one of the many NLP problems that need to be solved to achieve human-like performance in sentiment analysis.[9]

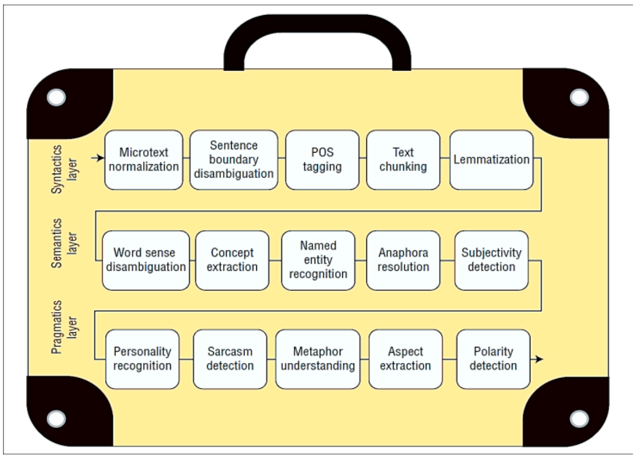

fig 2:Basic Process of Sentiment Analysis

It requires various steps and resolving numerous NLP problems to extract sentiment from a given text, such as negation handling and sarcasm handling. The basic process of sentiment analysis is as follows, as also shown in Figure 2. The data is first collected and extracted from various sources in diverse formats, which is then converted to text and processed using various NLP methods. The following steps are part of pre-processing :

1. Text pre-processing
2. Feature extraction
3. Feature selection

After the data has been pre-processed, this data is then passed as input to various types of classification algorithms based on a machine learning approach. At last, this processed data using various tools is represented and visualized in various formats like Bar Graph, Column Chart, Line Graph, Dual Axis Chart, Area Chart, Stacked Bar Graph, Pie Chart, Scatter Plot Chart, Bubble Chart, Waterfall Chart, Funnel Chart, Bullet Chart, Heat Map, etc.

Data extraction

The first step in sentiment analysis is to collect the raw data from various sources, which can be a single source or multiple sources on the internet. In this fast-paced world, about 2.5 quintillion bytes of data are produced every year. Some available data sources are listed below:

- *Social Media:* Social media is one of the important sources for collecting data because of its diverse nature, like textual data in the form of tweets and comments on posts and text from various images and videos. It is a perfect data source to have an outlook on modern society and its changing lifestyle and personality.[10]

- Review Websites: These kinds of websites are those where people write their reviews and opinions on numerous types of products, services, and businesses. It has a huge database in a categorized manner, from which selecting a product is easy. A few of these websites are famous e-commerce sites like “ https://www.amazon.com” or ”https://www.walmart.com”

- Blogs: blogs are simple websites that comprise short paragraphs of information and opinions about certain products and services. These blogs are generally like blog posts and arranged chronologically, the latest one being on the top along with the date and time, and the name of the publisher[11]

- Forums: forums are like a message board where like-minded people discuss various topics, ask a bunch of questions, and reply to the answers, which are like threads. As these discussions are about one entity, people write their own opinions, and this user-generated data becomes interesting to extract the sentiment using various techniques. The development of the internet, primarily Web 2.0, has made collecting data easier for sentiment analysis and in various formats like plain text, CSV, XML, HTML, and various media-rich files.

A. Data Pre-processing

i. Text Preprocessing and Feature Extraction

The data that we receive from the above-mentioned sources is often noisy, wrong, or inconsistent data, or can be either way much too data or can be missing data, which becomes essential to be cleaned and preprocessed before passing it on to the various sentiment analysis algorithms.[12] Such dirty data, if not cleaned, can reproduce wrong conclusions and

DOI: 10.17148/IJIREEICE.2025.13831

bad models. Such data can also exist in a variety of types, such as mixed and unstructured data, and is to be processed into a structured or ordered data form. The various methods for converting noisy high-dimensional data to low-dimensional data are as follows:

Tokenization: Tokenization is the process of converting the given text into a smaller form known as tokens. This ensures that unwanted tokens are filtered out. A document will be converted into paragraphs, which then will be converted into sentences and eventually words.

Removal of stop words: There are various words that get repeated several times, which do not contain any context or convey any deeper meaning for the sentiment analysis. Examples of such words are “the”, “for”, “such”, etcetera.

Part-of-Speech (Post) tagging: This step is used for analyzing various structural elements of text, such as verbs, nouns, adjectives, and adverbs.

Normalization: The words in a text could have been written in a mixture of cases, which needs to be converted into one case so that all words can be treated equally. The same word, such as “ text”, could be written in either “text”,” TEXT”, or “Text”, which all have the same meaning but can require several lines of code to handle. So, normalization is an

An essential tool to save a lot of resources and complexity. Another example is numbers written with the text “ I have 5 dogs”, which can be converted into “I have five dogs”.

Negation handling: It is the technique of converting apostrophes connecting words into two simple words with the same meaning so that these can be in standard lexicons. For example, “ don`t “ will be converted into “do not”.

Removing unwanted characters: standalone punctuations, special characters, and numerical tokens need to be removed as they do not contribute to any sentiment analysis.

Lemmatization: This normalization is very similar to stemming, but it accounts for the context of the word. It is a process of finding the base or dictionary form of a word, known as the lemma. For example, the words like ‘are’,’ is’, and’ being’ will be transformed into ‘be’

ii. Feature selection

The characteristics of a given data can be described by its features, which can be relevant, irrelevant, or completely redundant, and such unwanted features can be identified and removed by making use of various feature selection methods to reduce the size of the feature dimension space to improve the accuracy of sentiment analysis.

There are two ways of performing such a task:

1. Lexicon-based approach
2. Statistical methods

The lexicon-based approach is which is accomplished by humans by collecting terms having powerful sentiments, which can be used to build a small feature set, which then is enriched with other terms of a similar kind, which is a tedious process but quite effective.

The statistical method, on the other hand, is fully automated and widely used but is not very effective. These approaches are further classified as such:

1. Filter approach: This approach is the most common and computationally less expensive, suitable for datasets containing huge numbers of features. This approach does not use any machine learning algorithms and just relies on statistical measures and select the high-ranked features.[13] Some of these features are
   - Information gain
   - Chi-square
   - Document frequency
   - Mutual information

2. Wrapper approach: This type of machine learning technique is much more effective, but it is computationally intensive in finding the results. It works on the principle of evaluating a subset of features based on the previously applied machine learning approaches` results. It is the combination of machine learning algorithms and a feature subset generation strategy.[14]

3. Embedded approach: This approach uses algorithms which has the ability to select features of a given text to extract the sentiment on its own during its execution, which helps to eliminate the human need to select the important features. This is a better and more efficient approach than using the wrapper approach. Some of the algorithms based on this approach are CART, C4.5, ID3 [15]

4. Hybrid approach: This type of approach is much more effective than the above-mentioned approaches as it yields the best results in determining the sentiment of a given piece of text. It is the combination of the best features and functionality of filter and wrapper approaches, and thus is the choice of many companies and data scientists.[16]

DOI: 10.17148/IJIREEICE.2025.13831

IV. SENTIMENT ANALYSIS TECHNIQUES

After getting the raw data preprocessed for sentiment analysis, we get to the main method of getting the results by applying various methods and algorithms. What type of algorithms to use depends on our problems and the type of data we need to do our analysis on. Most of the literature divides sentiment analysis techniques into various approaches, which can give different types of results depending on the application and domain. Thus, this section will discuss the various approaches for performing sentiment analysis. [17]

A. Supervised learning

Supervised machine learning, as the name implies, requires some kind of supervision. In this case, we have to give certain labels to each of the features and then train them and test them with fresh data to check their accuracy and repeat the process until we get the required results. The labels here are generally positive, negative, and neutral.[19] Here we have four supervised classification approaches such as shown in the following figure 3.

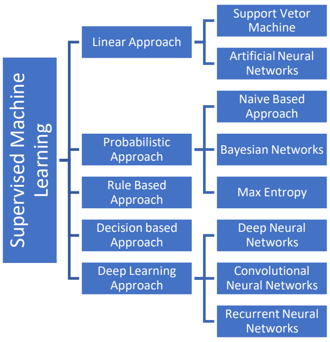

fig 3:Classification Approaches to Sentiment Analysis

a) Linear approach

A linear approach is based on pure statistics, which uses the concept of linear regression for classifying the analysis. It makes use of linear or hyperplane decision boundaries. Linear is used when using one class and a hyperplane is used for two or more classes. This classification is done using a linear predictor P=A.X+b, where A is the vector of linear coefficients, X is the document frequency of the word, and b is the bias. Linear approaches are further classified into Support Vector Machine (SVM) and Artificial Neural Networks(ANN).[20]

i. Support Vector Machine:

A support vector machine (SVM) is a non-probability classifier that can be used to decompose data linearly or non-linearly and can handle both discrete and continuous variables. It has a strong theoretical basis and performs more accurate classification than most other algorithms in many applications. SVM is suitable for text classification, making it popular in sentiment classification. The main goal of the SVM classifier is to find the optimal hyperplane to separate the classes. Efficient decomposition means that the hyperplane has maximum amplitude from the nearest training point of either class because a larger amplitude reduces the generalization error of the classifier[21]

ii. Artificial Neural Network:

Artificial Neural Network (ANN) has become an important classification method and gained popularity in recent times. It is based on the idea of extracting features from linear combinations of provided data, such as an input and modeling the output as a nonlinear function of these features.
It mainly involves three layers: input, output, and hidden layers. A hidden layer can be a single layer or multiple depending on the application and algorithm. These layers have many organized neurons and these links have some weight attached to them, which is obtained from minimizing the global error function in a gradient descent training process.[22]

b) Probabilistic Approach

Unlike the linear approachoutputs the most likely class of a given input (belonging to a positive or negative class), the probability classifier predicts the probability distribution over a set of classes, and they are usually based on the Bayes theorem. The classifier uses mixed models to perform the classification, where each class is a part of the mixture. These types of classifiers are also known as generic classifiers because each component in the mixture is a composite model.

Probabilistic classifiers are easy to implement, computationally fast compared to other algorithms, and they do not require much training data. However, the classification performance is sometimes worse if the data does not (at least almost) meet the distribution assumptions.[23]

i. Naïve Bayes

Naïve Bayes (NB) is a simple classifier and it is one of the most commonly used algorithms in the field of text classification. The model is based on Bayes' theorem and depends on BoW feature extraction. Therefore, the position of a word in the document is ignored and the presence of a particular word does not depend on the presence of any other words. Naïve Bayes assigns a document to the category c that maximizes P(c|d) by applying Bayes' rule :

$$\mathrm{P}(\mathrm{c}|\mathrm{d}) = \frac{\mathrm{p}(\mathrm{c})\mathrm{p}(\mathrm{c}|\mathrm{d})}{\mathrm{p}(\mathrm{d})}$$

where p(c) is the prior probability of category c, p(d|c)is the prior probability of document being assigned to category c, and p(d)is the prior probability of document d.[24]

ii. Bayesian Network

The Bayesian network (BN) consists of a directed acyclic graph where each node represents a random variable and the edges between the nodes represent the influence relationship. The model assumes that all nodes are independent because they are random, while on the other hand assumes that these nodes are completely dependent because of the conditional dependencies between them. It is a complete architecture for describing relationships between a set of variables through a common probability distribution, and because of its extensible structure, it is easy to add new variables. In the field of text classification, Bayesian networks help find relationships between a large number of words.[25]

iii. Max Entropy

Maximum Entropy (ME), also known as Conditional Exponential Classifier or Maximum Classifier, makes no assumptions about relationships between features. It estimates the conditional distribution of class c labels for a document d to maximize the entropy of the system using the following exponential form.

$$\mathrm{P}_{\mathrm{ME}}\,(\mathrm{C}|\mathrm{D}) = \frac{1}{\mathrm{Z}(\mathrm{d})} \exp\,(\sum_{\mathrm{i}} \lambda_{\mathrm{i}\,,\ \mathrm{c}} \quad \mathrm{f}_{\mathrm{i}\,,\ \mathrm{c}}(\mathrm{d},\mathrm{c}))$$

Where Z (d) is a normalization function, fi,c is a feature function for the feature fi and the class c, and λi,c is a parameter for the feature weight to make sure that the observed features match the expected features in the given set.[26]

$$\mathrm{f}_{\mathrm{I,c}}(\mathrm{d},\mathrm{c}') = \begin{cases} 1, \mathrm{n}_{\mathrm{i}}\ (\mathrm{d}) > 0 \ and\ c' = c \\ \quad 0, \mathrm{otherwise} \end{cases}$$

c) Rule-Based Approach

The rule-based approach. The term rule-based classifier can be used to refer to any classification scheme that uses IF-THEN rules for class prediction. Therefore, the classifiers involved in this technique depend on a set of rules to perform sentiment classification. A rule can be expressed as LHS → RHS, where the left side (LHS) represents a precursor to the rule or a set of conditions on the feature set represented in the DNF (Disjunctive Normal Form).) and the right side (RHS) represents a conclusion or consequence (class label) of the rule if the LHS is satisfied. The rule-based classifier can quickly classify new instances, and its performance is comparable to a decision tree. Another advantage of the rule-based approach is that it can avoid over-equipping. However, interpreting them becomes difficult and extremely laborious if there are too many rules. Furthermore, it has poor performance against noisy data.[27]

d) Decision tree approach.

In this approach, the training data space is hierarchically decomposed using an attribute value condition to classify the input data into a finite number of predefined classes. The condition for attribute values is with or without one or more words. This tree approach is a flowchart-like structure in which each inner node specifies a test on an attribute, each branch represents a result of the test, and the leaf nodes represent child nodes or class distributions. Decision tree classifiers are easy to understand and understand, and they can handle noisy data. However, on the other hand, they are unstable and prone to being over-equipped. The decision tree approach performs very well on large datasets and hence it is not advisable for small datasets.[28]

e) Deep Learning

The application of ANN-based deep learning (DL) for emotion analysis has become very popular recently. DL is an emerging field of machine learning that provides methods of representing learning features in a supervised or unsupervised manner. The term "deep learning" refers to the multi-layered perceptron neural network inspired by our brains. Thus, with this architecture, it is possible to train more complex models on a much larger data set, and thus produce state-of-the-art results in many application areas, from machine vision computing and speech recognition to NLP. DL includes many neural network models such as CNN (Convolutional Neural Network), RNN (Recurrent Neural Network) and DBN (Deep Belief Network). These models do not need to come with predefined features selected by the engineer, but they can learn complex features on their own from the data set. On the other hand, they are complex and very time-consuming to compute. Several studies have addressed deep learning approaches to sentiment analysis in detail. However, the following subsections briefly and briefly describe the most common deep learning models used for sentiment analysis.[29]

i. Deep neural Networks(DNN)

This model is an artificial neural network (ANN) with many layers (hidden layers) between the input and output layers. The input layer consists of input data, the hidden layers consist of processing nodes called neurons, and the output layer consists of one or more neurons used to generate the network output. It uses complex mathematical modeling and an ANN's learning ability to find the right relationship, whether linear or non-linear, to map input to output. The flow process of ANN and certainly DNN can be classified as downstream and downstream. The next ANNs are simple networks and are therefore suitable for sentiment classification. DN Architecture and its variants (e.g., CNN and RNN) have been used in many NLP tasks, including sentiment analysis. Vassilev designed a model called BowTie based on a deep feedback neural network, which consists of an encoding layer, a layer of hidden layers, and an output layer. The evaluation of this model shows promising results compared with other methods.[30]

ii. Convolutional neural networks (CNN)

This architecture is a special type of feedforward neural network that was originally used in the field of computer vision, but recently it has achieved good results in various fields such as recommendation systems and NLP.. CNN's layers consist of an input layer, an output layer, and a hidden layer consisting of multiple composite layers, grouping layers, normalized layers, and fully connected layers. Transformation classes filter the input (e.g., integrating words into a text sentiment classifier) to extract features, while clustering classes reduce feature resolution to make feature detection independent of noise and small noise changes. The normalized layer normalizes the output of the previous layer to improve convergence during training, and fully connected classes are used to perform the classification task. CNN has recently become very popular in the field of sentiment analysis.[31]

iii. Recurrent neural networks (RNN)

This model uses a memory cell to process a sequence of inputs. The ability to capture and store information over a long sequence makes RNNs widely used in NLP tasks such as sentiment analysis. In an RNN, the output depends on all previous computations. For example, to predict the next word in a sentence, the model uses all the states of the previous words and the relationships between them. One of the main problems of standard RNNs is the leakage gradient, and to overcome this problem, Hochreiter and Schmidhuber introduced a special type of RNN called LongShort Term Memory (LSTM), which is becoming very popular.in many places. This architecture is increasingly used by many researchers to classify sensations.[32]

B. Unsupervised Learning

Most of the existing methods for sentiment analysis are based on supervised learning models that are trained from a labeled repository, where each document was labeled before training. But sometimes it is difficult to collect and generate labeled data sets, especially for mostly unstructured text data. That's because their generation asks people to label data is too laborious and time consuming. On the other hand, it is easier to collect unlabeled datasets than, classify them using unsupervised learning methods. These techniques use the statistical support of document errors, such as word co-occurrence, NLP processes, and vocabulary with emotionally polarizing words. However, in machine learning, unsupervised approaches in the field of sentiment analysis often use grouping, which can classify data into different categories without specifying exactly which type of emotions are expressed in each category. In other words, the Clustering method divides data into groups (clusters), where data in a cluster are more similar from a particular point of view than data from different clusters. [33]

Unsupervised Learning can be categorized into two, as shown below in fig4.

DOI: 10.17148/IJIREEICE.2025.13831

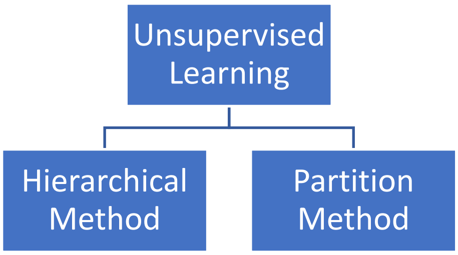

fig 4:Unsupervised Learning categories of Sentiment Analysis

a) Hierarchical Method

Hierarchical methods produce a hierarchical decomposition of a data set represented by nested clusters (groups with subgroups) organized as a tree. Hierarchical techniques can be divided into two main strategies, i.e., agglomeration and clustering. Split clustering is known as a top-down approach. This method starts with an individual cluster that groups all the data, then assigns this data to sub-clusters through a recursive process based on their similarity. Tsagkalidou et al. used this approach to propose a framework for grouping blog posts based on how close they are to certain emotions. Aggregate clustering (also known as the bottom-up approach) considers each piece of data to start in its own cluster and then merges clusters containing similar data until one or a set is left. Archambault et al. applied cumulative clustering to discover themes and sentiments in microblogging data.[34]

b) Partition methods

Partitioning methods are designed to partition data into discrete sets of clusters, where each element is assigned to only one cluster. This division is based on a similarity criterion, which is usually the Euclidean distance between elements. The data within the cluster have a very short distance from each other, but the greatest distance is from the data in other clusters. The most widely used segmentation algorithm is the k-means algorithm and its variants. The K-means algorithm iteratively assigns data objects in a data set to cluster centroids based on the similarity between the data objects and cluster centroids, starting with a predetermined number of initial cluster centroids. The process stops when the convergence criterion is met. The criterion may be a fixed number of iterations, or the result may not change after a certain number of iterations.[35]

C. Semi-supervised Learning

Semi-supervised learning (SSL) approaches are also used when there are difficulties in obtaining labeled data, but unlike unsupervised approaches, this technique uses a small set of initially labeled training data to guide the feature learning procedure. Thus, it fits in between supervised and unsupervised approaches. SSL approaches make full use of large amounts of low-cost unlabeled data, save a lot of time and effort, and gain a classifier with strong generalization ability in addition to more labeled data. Hussain and Cambria proposed a novel semi-supervised learning model for Big Social Data analysis. It is based on the combined use of random projection scaling and SVM. The results show that these semi-management models can significantly improve the performance of some NLP tasks, including sentiment analysis. Recent research on SSL-based sentiment analysis can be divided into five categories: generative, collaborative learning, self-learning, graph-based learning, and multi-representation learning. [36]

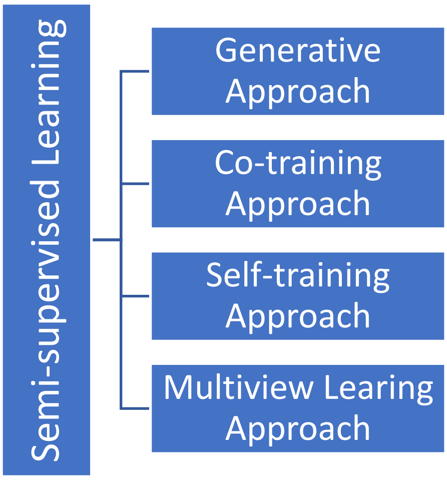

fig 5:Semi-supervised Approaches to Sentiment Analysis

DOI: 10.17148/IJIREEICE.2025.13831

a) Generative Approach

This approach assumes that the parameters of each distribution can be estimated if data from different categories follow different distributions and there is at least one labeled information for each category. In other words, a generative model predicts the label (class) of the test input after determining the distribution over the input, training this model for each class using Bayes' rule. Menil et al. [37]proposed a very simple and powerful ensemble system for sentiment analysis that combines a complementary tree and a conceptually basic model. One of them was based on a generative approach. Overall, the system achieves a new level of performance on the IMDB movie review dataset. This proves that ensemble learning can also be used as a semi-supervised or unsupervised approach.

b) Co-training Approach

This algorithm was originally developed by Blum and Mitchell and assumes that data can be represented using two independent views, where each view has information about each data point. In co-training, two separate classifiers will be trained to teach each other based on the shared information between them during the training process. Each classifier was trained on a different feature set corresponding to the two views of the data. The training process is iterative, and at each iteration, co-training updates the dataset by adding the most confident classified instances from each classifier to the labeled data. The process stops when all unlabeled data have been used or a specific number of iterations has been reached. This algorithm has been used for sentiment analysis in several studies. In the work of Xia et al., the Blum and Mitchell algorithm [38]was used to propose a two-representation cooperative learning approach that solves the negation problem and improves the bootstrap efficiency for semi-supervised sentiment classification.

c) Self-training Approach

This approach is commonly used in semi-supervised learning. The learning process is divided into two stages. In the first step, the classifier is trained on a small amount of labeled data. In the second step, the trained classifier is used to classify unlabeled data, adding the most reliable samples to the original training set as new labeled data. The last step is iteratively repeated with the new labeled data. The resulting model is then evaluated using the test data. This approach has been widely used in the field of sentiment analysis. He and Zhou proposed a new framework based on a self-learning approach that learns from labeled features rather than labeled instances. Experimental results showed that their approach outperformed some existing methods.[39]

d) Multi-view Learning Approach

This approach involves considering multiple perspectives on the problem, and the overall outcome is achieved through consensus. Each classifier is trained on one representation and then uses these classifiers to label unlabeled samples, adding them to the training set if they are classified with high confidence. This method is usually applied to jobs with multiple feature sets. Lazarov and Koychev propose an approach based on multi-viewpoint learning to analyze the sentiments of movies in the Bulgarian language.[40]

V. APPLICATIONS OF SENTIMENT ANALYSIS

With the introduction of Web 3.0, the world is seeing a substantial growth in the use of the internet for their jobs, education, entertainment, security, for buying and selling products and services over the internet. This emergence has led many companies to use many latest and growing technologies to cater to the public's needs and make a lot of profit in doing so. Here, sentiment analysis plays a huge role in making their task so easier, and this can be automated for better and faster results. Not only companies, but the government is also using sentiment analysis to make new policies and laws and monitor the public and ensure their safety. There are several sectors that are making use of sentiment analysis, which are listed below.

A. Commercial Sector

Today, all of the businesses and companies are making use of sentiment analysis along with data science and statistics and better marketing strategies to improve their products and services. People generally like to give their reviews and comments on the shopping sites and companies` own websites for allowing customers to write about the products and services. People can rate and comment on the various aspects and features of the products, which can be seen by different people and they can give their vote and express their sentiments and opinions by giving stars or equivalent likes. All of this huge data can be interpreted and the above-mentioned techniques can be applied to get results. These results are then used by companies to make certain changes in their upcoming products or improve their services to make a profit and stay ahead of the competition and also improve brand reputation and sales.

DOI: 10.17148/IJIREEICE.2025.13831

Sentiment analysis here is helping a company to get vital information about their product, opinion of customers and their sentiments on the products, whether they like or dislike the whole product or certain aspects of it, so that they can make it better and ditch the unwanted features and services, and this cycle continues. Along with this, companies can improve their marketing strategy by doing market research and analysis, which can bring in more customers and can make their loyal customers well satisfied, which makes both companies and investors make huge profits.
Using such technology, customers have their choice of comparing various products on various E-Commerce websites and apps, on the basis of their individual features, and make better purchasing decisions, and stay away from bad products which is hated by hundreds or thousands of customers.

B. Recommendation system

Various companies, such as E-commerce websites like “Amazon”, “Flipkart”, “Walmart” and video and music streaming services like “Netflix”, “Amazon Prime”, “YouTube”, “Hulu”, “Spotify”, and more, use a sentiment-based recommendation system for recommending their products and services. For both types of service providers, customers search for the product which have distinguishing characteristics and those can be used by these companies to push the same type of product at a cheaper price or products based on those searched prices. E-commerce sites recommend these products to lure in customers into buying more products, much more than they want, by presenting them as bundled or exclusive deals, tricking customers as if they were to be saving cash on such combo deals.

Video and music streaming sites also do the same when a subscriber watches or listens to music of a particular genre, the recommendation system pushes them recommended videos, shows, movies, or songs, either based on that genre or artist, or performer. These systems map out a unique identification token by analyzing certain features of that particular product or service and make a group of individuals for either serving them unique content or products they like, or serving them advertisements based on their preferences.

The revenue generated by use of recommendation system is huge as they kind of trap their users in and use their services by either making them buying more of the products or services, or In case of streaming services generate profit by serving them personalized ads or letting their customers pay them money to use their services and now a days these steaming services have certain paid tiers such as “Basic”, “Premium”, “Student Plan”, “Family Plan” and “Ultra”, which all have different pricing and benefits for certain amount of time and for certain devices and members. Customers tend to enjoy the content and keep paying these companies the money to get themselves entertained.

C. Government Sector

Besides the companies using the sentiment analysis, governments across the world use it for keeping a check on the public and making rules and regulations and various policies for improvement of the country`s economy and stability of government, betterment of the people and ensuring their safety and keeping crimes and unlawful activities in check. Twitter has seen tremendous growth in the past few years and people are using it to tweet and post about their social life. The government is making use of sentiment analysis to know about the various activities related to politics, religion and various social activities to keep an eye on the public and ensure there are no nefarious activities. Moreover, the government can make use of such social media platforms to get public opinion on the new policies and see the public reaction, whether they like it or not, or are just neutral about it and certain changes they can make to improve on it. The government can scrape certain websites that have blogs, articles, and forums about various public views on the working of the government and its already implemented policies. Such use of analysis is being used to categorize various groups of individuals who can be a threat to the nation if they tend to post or tweet online, which can result in violence or acts of terrorism. Such technologies have been used during election campaigns to analyze which political party has a high probability of winning the election by just analyzing the Twitter hashtags related to elections and eliminating the unlawful practices used by the party to manipulate the public to make the election one-sided.

So, there are various benefits of using sentiment analysis for various companies, the government and researchers for human advancement and a better internet and society. Many researchers are working on this particular field to find more benefits and domains to bring profit to companies and service to mankind.

VI. CHALLENGES TO SENTIMENT ANALYSIS

Although the sentiment analysis technology has gotten better over the years with great results and various applications in different domains, it still has to face many challenges, as our human language isn’t as easy and full of errors and slang and hundreds of dialects for a particular language. Here are some of the major challenges that need to be tackled to make this technology better and versatile.

A. Sarcasm detection

Sarcasm is an activity of saying or writing the opposite of what someone means, which is done to make another person feel stupid or show them that one is angry. We often use sarcasm a lot in our day-to-day activities and analyzing this through a machine learning algorithm becomes a tedious task. This problem of identifying text whether it has a positive or negative tone when one has spoken or written in a positive tone, but to a listener or reader it may sound negative or vice versa. While doing sentiment analysis can become a great problem, which can literally alter our results. The complexity and ambiguity that could arise due to the use of a sarcastic tone is one of the hurdles that researchers are working to solve and get accurate results during analysis.

B. Negation Handling

There are certain day-to-day use words like "not, "neither", "nor", etc., which are very essential for sentiment analysis, which can alter the meaning of the sentence. But there are many sentiment analysis approaches that tend to remove these negation words because they are part of the Stop-Word list, as these words represent the neutral sentiment in a lexicon which do not have any impact on final polarity. But getting these words removed during analysis changes the results and does not represent the true emotion, which defeats the true purpose of analysis. For example, "This is a nice car." This sentence is self-explanatory and so is this one: "This is not a nice car". But if during pre-processing the word "not" is removed, both sentences become the same when it is clearly not.

C. Spam Detection

Spam is undesired content that can be automatically created by various online bots or by humans who provide the wrong or misleading information, which is used by hackers or nefarious people either to steal money, sensitive information, or alter the mindset of people to do tasks they wouldn't otherwise do. Spam is increasing at a high speed on the internet. We either get it in our email inbox, or it can be in the form of fake comments or reviews. These spam comments or reviews, when read by people, can change the perception of a product or a service, which can either sell or represent a bad product by making it look good or vice versa. Dealing with such spam content is hard, as these are just normal words that tend to have a normal meaning, but are bad for a product or company`s reputation. This spam can be spread by both companies for selling a bad product by luring a customer with the reviews on the product page, or it can be done by someone else to defame a company. So, it is essential to remove spam from our analysis to get appropriate results.

D. Low-Resource Language

This paper has dealt with the use of the English language for sentiment analysis for reasons that the English language is an international language, which means that international business and commerce are done using the English language, and most of the internet users across the world use English as their primary language. But certain parts of the world use languages other than English, which is a problem, as unlike English, other languages do not have many of the linguistic resources for sentiment analysis. Coding the resources for such languages is very time-consuming and is not very cost-effective, and converting them to English for analysis may alter the meaning of the text. These types of languages, which do not have many linguistic resources, are referred to as Low-Resource languages.

E. Code-mixed data or Multilingual data

Another big problem during sentiment analysis is the use of different languages and dialects, which is really difficult for the algorithm to handle. For example, Hinglish is a mixture of English and Hindi, which is used by Indians. Handling such multilingual languages is not an easy feat, which requires a lot of manual labor to separate the words from different languages. There is also the trend of using slang with this multilingual language, like Hinglish, which makes the task more difficult to analyze. The usage of vocabulary and syntax from different multiple languages known as Code-Mixing, which lacks the formal grammar is one of the problems that needs to be dealt during analysis because there are many important comments and reviews by millions of people over plethora of different sites, social media platforms and forums which cannot be ignored, hence there is need for building the system for handling such a problem.

F. Emojis

People tend to decorate their texts with appropriate symbols and emojis, which may seem cute to look at humans, but are a huge hurdle for an algorithm to get desired results from such emoji-filled characters. It may be okay for some texts where a correctly written sentence, such as "This is a lovely landscape🏔". If during analysis this emoji is removed, we can still get an accurate result. But people now a days, especially, on social media, blogs, forums, or even review sites, writes like this for example, "These 🏔 are very🥶❄, we need to 🔎👀 a ♨ place" which should look like this if actual words are to inserted in place of emojis "These mountains are very chilly, we need to look for a warm

DOI: 10.17148/IJIREEICE.2025.13831

place”. So the usage of such emojis in such places may look fun to watch, but difficult to get any satisfactory results out of it. So, we need to find a solution to tackle such problems.

VII. CONCLUSION

This paper presents an overview of sentiment analysis technology, its various approaches and the domains it is used in. This paper explains some of the most used methods for performing an analysis that is heavily focused on data mining and preprocessing and machine learning. After that, various categories of analysis were presented in detailed form so that the reader can get an idea of the technology. After that, various uses in different sectors have been discussed, followed by some of their challenges that need to be addressed. Sentiment analysis showed here focuses on the English language, but with the right tools and methodology, it can be used for various languages worldwide.

REFERENCES

[1]. B. Liu, Sentiment Analysis, 2015, pp. 1–367, https://doi.org/10.1017/CBO9781139084789
[2]. M.V. Mäntylä, D. Graziotin, M. Kuutila, The evolution of sentiment analysis—A review of research topics, venues, and top cited papers,
[3]. Comput. Sci. Rev. 27 (2018) 16–32, https://doi.org/10.1016/j.cosrev.2017.10.002.
[4]. H.H. Do, P. Prasad, A. Maag, A. Alsadoon, Deep learning for aspect-based sentiment analysis: A comparative review, Expert Syst. Appl. 118 (2019)272–299, https://doi.org/10.1016/j.eswa.2018.10.003.
[5]. C.C. Aggarwal, Machine learning for text, Mach. Learn. Text. (2018) 1–493,https://doi.org/10.1007/978-3-319-73531-3.
[6]. S. Behdenna, F. Barigou, G. Belalem, Sentiment analysis at document level, in: SmartCom 2016, 2016, pp. 159–168, https://doi.org/10.1007/978-981-10-3433-6_20.
[7]. N. Indurkhya, F.J. Damerau, Handbook of Natural Language Processing,2010, pp. 1–704.
[8]. B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 5 (2012) 1–167, https://doi.org/10.2200/S00416ED1V01Y201204HLT016.
[9]. O. Alqaryouti, N. Siyam, A.A. Monem, K. Shaalan, Aspect-based sentiment analysis using smart government review data, Appl. Comput. Informatics.(2019) 1–20,https://doi.org/10.1016/j.aci.2019.11.003
[10]. E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intell. Syst. 32 (2017) 74–80, https://doi.org/10.1109/MIS.2017.4531228.
[11]. B. Batrinca, P.C. Treleaven, Social media analytics: a survey of techniques, tools and platforms, AI Soc. 30 (2015) 89–116, https://doi.org/10.1007/s00146-014-0549-4.
[12]. Y.H. Gu, S.J. Yoo, Z. Jiang, Y.J. Lee, Z. Piao, H. Yin, S. Jeon, Sentiment analysis and visualization of Chinese tourism blogs and reviews, in:2018 Int. Conf. Electron. Information, Commun, IEEE, 2018, pp. 1–4, https://doi.org/10.23919/ELINFOCOM.2018.8330589.
[13]. B. Liu, Sentiment analysis and opinion mining, Synth. Lect.Hum. Lang. Technol. 5 (2012) 1–167, https://doi.org/10.2200/S00416ED1V01Y201204HLT016.
[14]. G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, F.E. Alsaadi, Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods, Appl. Soft Comput. 86 (2020)1–14, https://doi.org/10.1016/j.asoc.2019.105836.
[15]. R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: Introduction and review, J. Biomed. Inform. 85 (2018) 189–203,https://doi.org/10.1016/j.jbi.2018.07.014
[16]. H. Liu, M. Zhou, Q. Liu, An embedded feature selection method for imbalanced data classification, IEEE/CAA J. Autom. Sin. 6 (2019) 703–715,https://doi.org/10.1109/JAS.2019.1911447.
[17]. G. Ansari, T. Ahmad, M.N. Doja, Hybrid filter–wrapper feature selection method for sentiment classification, Arab. J. Sci. Eng. 44 (2019)9191–9208, https://doi.org/10.1007/s13369-019-04064-6
[18]. A. Collomb, L. Brunie, C. Costea, A study and comparison of sentiment analysis methods for reputation evaluation, in: Cogn. Informatics Soft Comput, 2013, pp. 1–10, https://liris.cnrs.fr/Documents/Liris-6508.pdf.
[19]. N.N. Yusof, A. Mohamed, S. Abdul-Rahman, Reviewing Classification Approaches in Sentiment Analysis, 2015, pp. 43–53, https://doi.org/10.1007/978-981-287-936-3_5.
[20]. H. Sankar, V. Subramaniyaswamy, Investigating sentiment analysis using a machine learning approach, in: 2017 Int. Conf. Intell. Sustain. Syst, IEEE,2017, pp. 87–92, https://doi.org/10.1109/ISS1.2017.8389293
[21]. A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of Tricks for Efficient Text Classification, 2016, pp. 1–5, http://arxiv.org/abs/1607.01759.
[22]. C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995)273–297, https://doi.org/10.1007/BF00994018

[23]. G. Vinodhini, R.M. Chandrasekaran, A comparative performance evaluation of neural network-based approach for sentiment classification of online reviews, J. King Saud Univ. - Comput. Inf. Sci. 28 (2016) 2–12,https://doi.org/10.1016/j.jksuci.2014.03.024.

[24]. D. Fisch, E. Kalkowski, B. Sick, Knowledge fusion for probabilistic generative classifiers with data mining applications, IEEE Trans. Knowl. DataEng. 26 (2014) 652–666, https://doi.org/10.1109/TKDE.2013.20.

[25]. .K.M.A. Hasan, M.S. Sabuj, Z. Afrin, Opinion mining using Naïve Bayes, in:2015 IEEE Int. WIE Conf. Electr. Comput. Eng, IEEE, 2015, pp. 511–514, https://doi.org/10.1109/WIECON-ECE.2015.7443981

[26]. L. Gutiérrez, J. Bekios-Calfa, B. Keith, A Review on Bayesian Networks for Sentiment Analysis, 2019, pp. 111–120, https://doi.org/10.1007/978-3-030-01171-0_10.

[27]. B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proc. 2002 Conf. Empir. Methods

[28]. Nat. Lang. Process. ({EMNLP} 2002), Association for Computational Linguistics, 2002, pp. 79–86, https://doi.org/10.3115/1118693.1118704.

[29]. A.K.H Tung, Rule-based classification, in: Encycl. Database Syst, Springer,US, Boston, MA, 2009, pp. 2459–2462, https://doi.org/10.1007/978-0-387-39940-_559.

[30]. J. Han, M. Kamber, J. Pei, 1 - Introduction, in: J. Han, M. Kamber, J. Pei(Eds.), Data Min, Third Ed., Third Edit, Morgan Kaufmann, Boston, 2012,pp. 1–38, https://doi.org/10.1016/B978-0-12-381479-1.00001-0.

[31]. L.M. Rojas-Barahona, Deep learning for sentiment analysis, Lang. Linguist.Compass. 10 (2016) 701–719, https://doi.org/10.1111/lnc3.12228.

[32]. J. Schmidhuber, Deep learning in neural networks: An overview, NeuralNetworks. 61 (2015) 85–117, https://doi.org/10.1016/j.neunet.2014.09.00.

X. Ouyang, P. Zhou, C.H. Li, L. Liu, Sentiment analysis using a convolutional neural network, in: 2015 IEEE Int. Conf. Comput. Inf. Technol. UbiquitousComput. Commun. Dependable, Auton. Secur. Comput. Pervasive Intell.Comput, IEEE, 2015, pp. 2359–2364, https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.349.

[33]. A. Aziz Sharfuddin, M. Nafis Tihami, M. Saiful Islam, A deep recurrent neural network with bilstm model for sentiment classification, in: 2018, Int. Conf. Bangla Speech Lang. Process, IEEE, 2018, pp. 1–4, https://doi.org/10.1109/ICBSLP.2018.8554396

[34]. Y. Han, Y. Liu, Z. Jin, Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers, Neural Comput.Appl. 32 (2020) 5117–5129, https://doi.org/10.1007/s00521-018-3958-3.

[35]. H. Suresh, S. Gladston Raj, A Fuzzy-Based Hybrid Hierarchical Clustering Model for Twitter Sentiment Analysis, 2017, pp. 384–397, https://doi.org/10.1007/978-981-10-6430-2_30.

[36]. X. Cui, T.E. Potok, P. Palathingal, Document clustering using particle swarm optimization, in: Proc. 2005 IEEE Swarm Intell. Symp. 2005. SIS 2005, IEEE, 2005, pp. 185–191, https://doi.org/10.1109/SIS.2005.1501621.

[37]. Y. Han, Y. Liu, Z. Jin, Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers, Neural Comput. Appl. 32 (2020) 5117–5129, https://doi.org/10.1007/s00521-018-3958-3.

[38]. G. Mesnil, T. Mikolov, M. Ranzato, Y. Bengio, Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews, 2014, pp. 1–5, http://arxiv.org/abs/1412.5335.

[39]. A. Blum, T. Mitchell, Combining labeled and unlabeled data with cotraining, in: Proc. Elev. Annu. Conf. Comput. Learn. Theory - COLT' 98,ACM Press, New York, New York, USA, 1998, pp. 92–100, https://doi.org/ 10.1145/279943.279962.

[40]. W. Gao, S. Li, Y. Xue, M. Wang, G. Zhou, Semi-Supervised Sentiment Classification with Self-Training on Feature Subspaces, 2014, pp. 231–239,https://doi.org/10.1007/978-3-319-14331-6_23.

[41]. G. Lazarova, I. Koychev, Semi-Supervised Multi-View Sentiment Analysis, 2015, pp. 181–190, https://doi.org/10.1007/978-3-319-24069-5_17

[42]. L. Wang et al., "Dynamic Bandwidth and Wavelength Allocation Scheme for Next-Generation Wavelength-Agile EPON", J. Optical Commun. Networking, vol. 9, no. 3, pp. 33-42, 2017.

[43]. M. P. McGarry, M. Reisslein and M. Maier, "Ethernet Passive Optical Network Architectures and Dynamic Bandwidth Allocation Algorithms", IEEE Commun. Surveys & Tutorials, vol. 10, no. 3, pp. 46-60, 2008.

[44]. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT”, IEEE Electron Device Lett., vol. 20, no. 2, pp. 569–571,1999.

[45]. S. Sutar et al., "D-PUF: An Intrinsically Reconfigurable Dram PUF for Device Authentication and Random Number Generation", ACM Trans. Embedded Computing Systems (TECS), vol. 17, no. 1, pp. 1-31, 2017.

[46]. O. El Mouaatamid, M. Lahmer and M. Belkasmi, "Internet of Things Security: Layered Classification of Attacks and Possible Countermeasures", Electronic J. Information Technology, vol. 4, no. 9, pp. 256-261,2016.

[47]. M. Banayeeanzade et al., "Generative vs. Discriminative: Rethinking the Meta-Continual Learning", Advances in Neural Information Processing Systems, vol. 34,no. 6,pp. 124-131, 2021.