ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

Volume 11, Issue 2, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

IMPACT FACTOR: 7.583

| www.ijarasem.com | ijarasem@gmail.com | +91-9940572462 |

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 2, March 2024 |

Classification of Sentiment Analysis Techniques: A Comprehensive Approach

Rahul Sagar[1], Sumit Dalal[2], Rohini Sharma[3], Sumiran[4]

Student, Department of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[1]

Assistant Professor, Department of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[2]

Assistant Professor and Corresponding Author, GPGCW, Rohtak, Haryana, India[3]

Assistant Professor, Department of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[4]

ABSTRACT: Sentiment analysis, sometimes referred to as opinion mining, is a potent technique for figuring out how the general public feels about different organizations, merchandise, or happenings. Sentiment analysis has grown in significance for companies, governments, and scholars as a result of the widespread use of social media and online reviews. This study offers a thorough analysis of sentiment analysis approaches, including everything from conventional approaches to cutting-edge deep learning techniques. We talk about the problems with sentiment analysis, like sarcasm, ambiguity, and context-dependency, and we look at how various methods deal with these problems. Furthermore, we investigate diverse uses of sentiment analysis in a range of fields, such as politics, marketing, healthcare, and customer service. This paper intends to assist academics and practitioners in choosing suitable approaches for sentiment analysis tasks by providing a brief summary of current techniques and their uses.

KEYWORDS: Sentiment Analysis, Natural Language Processing (NLP), Machine Learning, Deep Learning, Text Classification, Sentiment Lexicons, Feature Engineering, Social Media Analysis (SMA), Customer Feedback Analysis.

I. INTRODUCTION

"Sentiment analysis (SA) or opinions mining (OM)" is the term used to describe the analytical study of human behavior, viewpoints, and sentiments about an object. The goal of the Natural Language Processing (NLP) technique known as SA is to obtain attitudes and opinions from texts[1] . As determining whether a given text exhibits a positive or negative attitude is one of SA's goals, this task could be considered a text-classification problem [2]. However, although sentiment analysis seems like a straightforward process, it actually entails assessing other NLP subtasks, such as subjectivity recognition and mockery. These days, companies, governments, and organizations—as well as researchers—take sentiment analysis very seriously. There should have been multiple earlier papers recognizing SA, a subfield of NLP, despite the lengthy and illustrious history of utilizing public opinion in decision-making. SA is still in operation even if it has changed in the current day. This study will look at the problems, strategies, uses, and algorithms related to sentiment analysis. It gives the task flowcharts, graphs, and basic tables that show comparative data analysis. As far as we know, current surveys tend to ignore some SA techniques in favor of hybrid, lexicon-based, and machine learning (ML) approaches. This study aims to provide the conceptual framework and talk about the difficulties and uses of it.

II. PROCESS FLOW OF SENTIMENTS ANALYSIS

The general sentiment analysis procedure is shown in Figure 1. Data were gathered after being extracted in various forms from a variety of sources, converted to text, and then subjected to natural language processing techniques. Features like "text pre-processing, feature extraction, and feature selection" are specifically included in the processing step. Feature extraction (FE), which involves sorting through text input to uncover significant information that could be used to enhance the model's performance, is an essential step in sentiment classification. As previously mentioned, a feature describes an attribute of the data. A feature could be neither relevant nor redundant. Various approaches of feature selection (FS) are employed to remove redundant and superfluous attributes. FS is a strategy that entails

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 2, March 2024 |

identifying and eliminating superfluous and unwanted qualities from the feature list in order to reduce the size of the feature dimension space and improve the accuracy of emotion categorization.

Statistical methods and procedures based on lexicons are employed in feature selection. Humans build the features of lexicon-based approaches. The process is usually started by compiling expressions that evoke strong emotions in order to create a reduced feature set. The next stage is to use online resources or synonym identification to add additional phrases to this collection. One well-known example of this approach is the Senti-WordNet vocabulary. Statistical methods are typically classified into four categories: filter approach, wrapper strategy, embedding technique, and hybrid approach. The most popular feature selection method is the filter method. It selects features that aren't used by other machine learning approaches based on the general characteristics of the training data. The feature is rated using a variety of statistical parameters, and the features that receive the highest rankings are subsequently chosen [3].

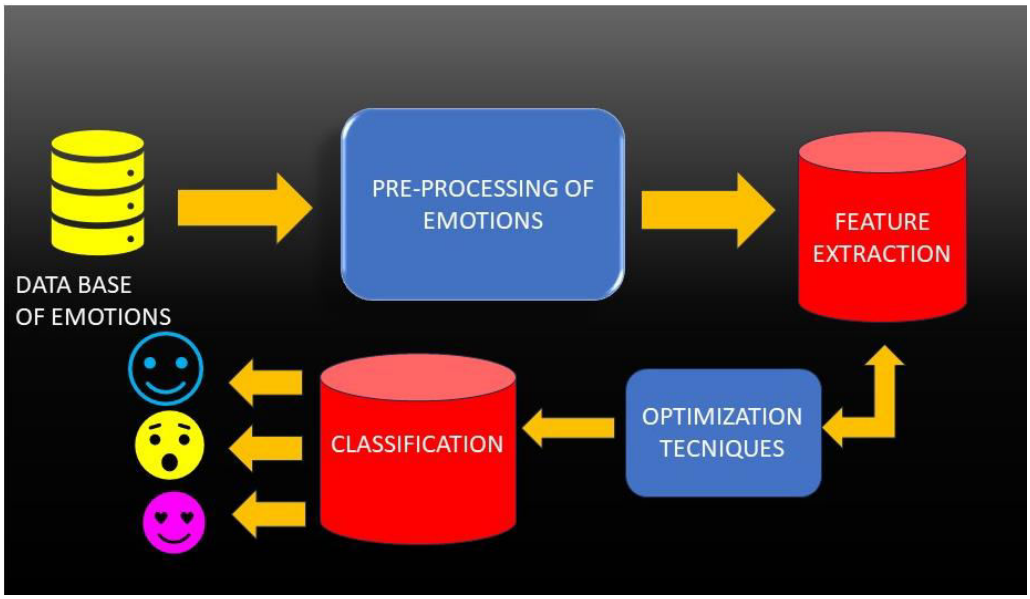


Figure 1: Processing of Data for Sentiments Analysis

III. CLASSIFICATION OF SENTIMENT ANALYSIS

Sentiment analysis is a dynamic, rapidly expanding subject of study with a wide range of applications. The goal is to improve sentiment performance analysis and deal with issues surrounding this subject. It is imperative to incorporate contemporary methodologies into sentiment research, taking into account divergent perspectives. However, the majority of studies frequently divide SA techniques into two categories: methods based on lexicons and machine learning. The lexicon-based approach utilizes the emotion lexicon, a set of terms that are frequently used to characterize positive or negative emotions.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 2, March 2024 |

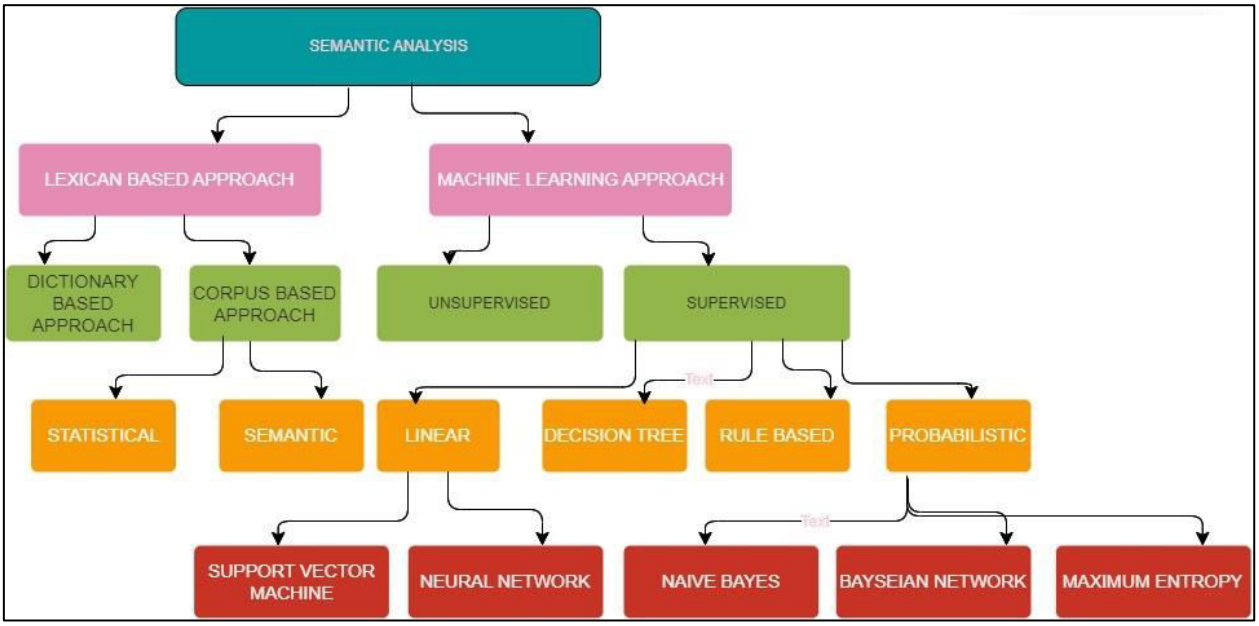


Figure 2: Categorization of semantic analysis approaches

LEXICON-BASED METHODS

Lexicons are collections of tokens with a fixed score that designates whether the material is favorable, unfavorable, or neutral. For a given review or paragraph, the Lexicon-based technique adds neutral, negative, and positive values to each token. Ultimately, an overall polarity is assigned to the statement based on the highest possible value of every person's judgment. Consequently, the text is divided into single-word tokens, and the polarity of each token is calculated and averaged. Particularly helpful for feature- and sentence-level sentiment evaluation is the lexicon-based approach. As there is no requirement for training data, this approach could be categorized as unsupervised.

The Lexicon-based approach [4] makes the assumption that a word or phrase's occurrence in the document correlates with sentiment since it presents raw text in a structured representation that has been processed. A lexicon is a collection of characteristics given an emotion value. It functions essentially as a dictionary, which is a predetermined list of terms with several synonyms that each word is connected to. Popular lexicons include SenticNet, WordNet, and others. The tweet additionally includes emoticons, which are emoji that represent sentiment and are among the lexicon's easiest to recognize in a tweet.

Dictionary-Based Method

This tactic is based on the notion that, in contrast to synonymous and antonymous terms, which have opposite emotional polarities, the sentiment lexicons in this approach are created using well-known dictionaries. Dictionary-based terms, like other dictionary-based approaches, suffer from the primary drawback of not being able to express conceptual words with precise phrases, which renders them inappropriate for dissemination and substance. In addition, writing dependence code takes a lot of effort and time.

To increase the size of the word set, certain terms are chosen as seed words in dictionary-based approaches, and these words are then utilized to find synonyms. Enlarge the size using online dictionaries. The unique and significant opinion words in a corpus are called seed words. In order to do sentiment analysis, these newly extended and seeded phrases are utilized. Dictionaries such as WordNet, SentiWordNet, SentiFul, and SenticNet are available. A dictionary-based approach to thesaurus lexicon creation has been suggested by authors [5]. He constructed a thesaurus using three online dictionaries, storing only terms that occur together in the lexicon to increase its reliability. The lexicon is expanded by the use of seed word synonyms and antonyms. Using a thesaurus increases the classification task's efficiency. TF-IDF techniques were used to choose the seed word. He also noted that this strategy requires a lot of time.

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 7.583 | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 11, Issue 2, March 2024 |

Corpus-Based method

A syntactic model, or a model akin to it, is used to describe the notion of tokens and their orientation in the bigger context after employing a list of words and their orientation. The model employs the syntactic and semantic structure to figure out the significance of a sentence. This method is distinct and effective for learning across several fields. For corpus-based methodologies, the following techniques are well-known. Using a corpus-based technique, we can identify a word's context orientation in addition to its label. Using this method, new subjective words are created from the corpus by first creating a list of seed words and then using their syntactic pattern. A word that appears together or with another word is said to have a syntactic pattern. The following describes statistical approaches and semantic techniques.

Statistical method: It is possible to find facts or patterns of relationship between concepts using statistical methods.

Semantic method: The comparative score of the markers for the emotional evaluation is computed using this technique. WordNet is frequently utilized for this purpose.

The authors conducted a theme analysis using sophisticated algorithms, which shortened the analysis's duration and increased its effectiveness [6]. Natural language processing (NLP) and text classification both benefit from the use of long short-term memory (LSTM). In this work, concordance lines from a corpus of news items were used to do a thematic analysis utilizing long short-term memory (LSTM) text categorization. To fully recognize the semantic shift of terms and concepts, yet, requires more than the statistical and quantitative analyses of corpus linguistics. In order to ascertain the degree of linguistic patterns that ought to be used in the experiment of the classification process, we thus propose that a corpus be classed from a linguistic conceptual point of view. Since this has been a drawback for many studies, we recommend looking into the concordance lines of the articles rather than just the link between collocates.

MACHINE LEARNING METHODS

ML approaches are used to classify sentiment polarity based on train and test datasets. These techniques can be divided into three categories, per [33]: "supervised learning," "semi-supervised learning," and "unsupervised learning." The supervised strategy is used when a classification program has a predefined set of classes; on the other hand, the unsupervised method could be important if it is hard to find this set since there isn't enough labeled data. On the other hand, unlabeled datasets containing a small number of tagged examples can be processed using the semi-supervised methodology. Reinforcement learning algorithms use trial-and-error strategies to maximize cumulative rewards and facilitate the agent's integration with its surroundings. Before being used with actual data, these algorithms need to be trained. Features can be extracted from text-formatted data. Word usage, profanity, and informative content can all be detected by SA systems that have been trained to look beyond basic content. Typical algorithms comprise the following:

Naive Bayes (NB)

NB method is used for both training and categorization. NB, a Bayesian classification algorithm based on Bayes rules, is applied. NB is a probabilistic classifier that uses the Bayes rule to predict the probability that a given set of features will be included with a given label. NB is frequently employed when the percentage of data used for training is low. Compared to negative NB categorization, positive NB categorization was 10% more accurate. After it was put into practice, average accuracy increased.

Support Vector Machine (SVM)

That method, which makes use of hyper-planes, analyzes data using the SVM approach to determine decision boundaries. Classification issues are a common application for SVMs, a type of supervised learning technique that is non-probabilistic. The basic aim of SVM is to find the hyperplane that most efficiently splits the data into subgroups. For this reason, SVM seeks the hyperplane with the largest possible margin.

With an F1 score of 83.4% and a mean AUC of 0.896, the linear SVM classifier was heavily applied [7]. Additionally, their technology demonstrated an improvement in email correspondence that was forecasted based on the sentiment of unseen emails. Unlike the traditional binary classification, the problem highlighted the need to assess the subjectivity of viewpoint and the expresser's legitimacy. Logistic regression is a machine training method wherein an input value is multiplied by a weight value. The classifier is the one that creates an awareness of the input features needed to

distinguish between positive and negative classes. It is possible to have any kind of independent variable. The predictive variables show minimal to no multicollinearity, and the dependent variable is binary, based to the LR approach.

K-Nearest Neighbors (KNN)

Sentiment analysis does not usually use KNN, but when trained correctly, the technique has been shown to produce good results. You can select the K value by any hyper-parameter tuning method, including grid search and randomized search cross-validation. The K nearest neighbor scores can be used to determine the polarity through either hard voting or soft addition.

Decision Tree (DT)

To classify the text's polarity, the supervised learning technique known as the DT Classifier builds a tree using the training example. These are used to determine which characteristic should be at the bottom and best represents reality. A graph-based reporting approach was presented to integrate sentence-level and sentence-level data [8].

Bayesian Network

While the Bayesian network can identify a semantic association between words, the Naïve Bayes classifier treats each word independently. The Bayesian network gives careful consideration to the interdependence of words. The Bayesian network uses an acyclic directed graph to describe dependency, with each node standing in for a word and the edges signifying the relationship between the variables. Authors [9] employed Bayesian networks as a sentiment classifier and discovered competitive output that was occasionally higher than that of other classifiers.

Unsupervised Learning

When labeled data reliability is a challenge, this approach is employed. Gathering unlabeled data is simpler than labeled data. The keyword lists of each category are used to classify the sentence. The unsupervised approach makes it easier to evaluate the domain-dependent data [10] [11].

Semantic analysis cab be further classified on following bases:

Subjectivity classification

Text fragments are categorized according to their subjectivity into objective and subjective (or opinionated) categories. Neutral information and facts are combined to make an objective sentence. After being adopted by three different families, split at birth, and born matching triplets, three strangers are reunited by an incredible coincidence. A subjective statement, on the other hand, conveys an individual's mindset, emotions, assessment, conviction, and more:

Polarity classification

Using polarity categorization, opinionated texts can be further separated into positive and negative categories. This method is useful for extensive analyses of both positive and negative trends found in text data, such as consumer feedback, social media posts, and reviews of goods. Beyond simple binary classification, advanced models also determine the sentiment strength. Text segments are divided into more than two categories in this instance; these groupings include extremely positive, positive, neutral, negative, and extremely negative. Not only is it possible to solve polarity and subjectivity tasks at the same time with the multiclass technique. Additionally, it yields more accurate results when processing sentences that contain modifiers (too, so, fully, etc.) and comparison expressions (better, most awful, etc.). However, there are numerous drawbacks to even multiclass polarity classification. It is unable to identify how customers feel about certain features or parts of your services, which is crucial for successful product creation and enhancement. That's where you might benefit from an aspect-based sentiment analysis.

Aspect-based sentiment analysis

The multi-step technique known as "aspect-based" or "feature-based" sentiment analysis aims to identify and retrieve sentiments regarding a particular element of a good or service. A single hotel review, for instance, may include both good and negative remarks. An aspect-based mechanism initially detects aspects discussed in the reviews or comments in order to gather all this important data. The text segments that mention certain aspects are then subjected to polarity

categorization. To determine the trending attitude toward a certain feature, the findings are finally combined and scored according to several aspects.

Fine-grained sentiment analysis

In this version, the sentiment target (subject) and its intensity are defined at the phrase or clause level using fine-grained analysis. It resolves issues that more basic approaches are unable to, such as handling comparative expressions.

IV. CONCLUSIONS

This review paper discusses the fundamentals of sentiment analysis and categorization techniques. The many sentiment analysis techniques and their performance metrics have been investigated in the systematic review. It has been noted that the quality of the chosen features and the classification algorithm depend on high classification accuracy. Much work has been done recently to find the semantic relationship using word embedding techniques and to classify using artificial neural networks. In addition to SA, Deep Learning (DL) techniques are widely applied in machine vision. Emoji-based SA is a good fit for this technology because of its innate capacity to automatically extract features and learn how to make sense of high-dimensionality data. Researchers are already using deep learning techniques like CNN and LSTM extensively, therefore in the future, hybrid or individual deep learning techniques may be utilized for evaluating sentiment in global contexts. In the machine learning technique comparison, the SVM has the highest accuracy (94.05%), while the Decision Tree (DT) has the highest precision rate (88.86%) and the SVM has the highest recall rate (87.34%) when compared to all other techniques.

REFERENCES

[1] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021.

[2] Y. Choi and H. Lee, "Data properties and the performance of sentiment classification for electronic commerce applications," *Inf. Syst. Front.*, vol. 19, pp. 993–1012, 2017.

[3] G. Adomavicius and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 896–911, 2011.

[4] Ł. Augustyniak, P. Szymański, T. Kajdanowicz, and W. Tuligłowicz, "Comprehensive study on lexicon-based ensemble classification sentiment analysis," *Entropy*, vol. 18, no. 1, p. 4, 2015.

[5] S. Park and Y. Kim, "Building thesaurus lexicon using dictionary-based approach for sentiment classification," in *2016 IEEE 14th international conference on software engineering research, management and applications (SERA)*, 2016, pp. 39–44.

[6] Y. Altameemi and M. Altamimi, "Thematic Analysis: A Corpus-Based Method for Understanding Themes/Topics of a Corpus through a Classification Process Using Long Short-Term Memory (LSTM)," *Appl. Sci.*, vol. 13, no. 5, p. 3308, 2023.

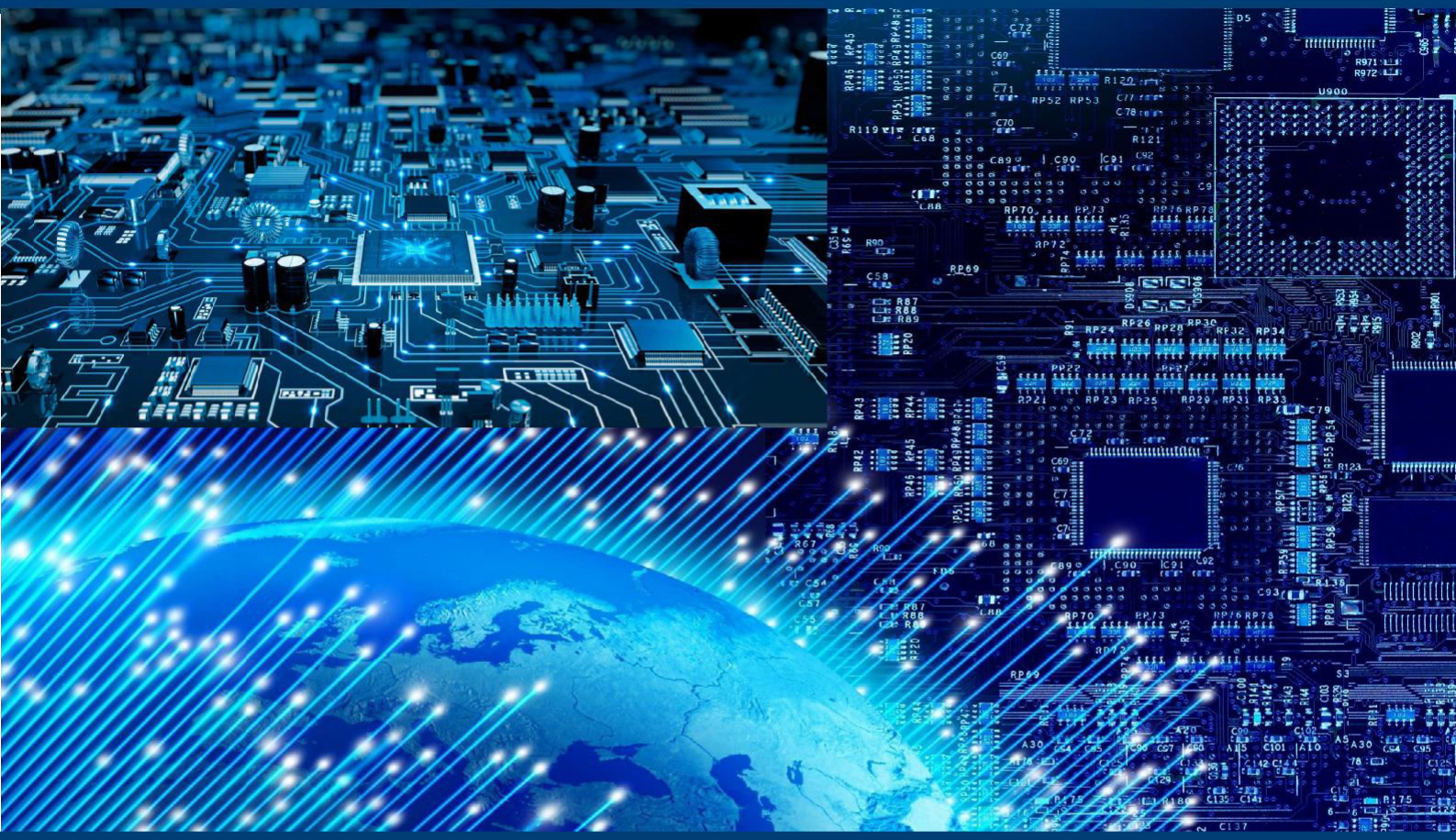
[7] F. Li, W. Wang, J. Xu, J. Yi, and Q. Wang, "Comparative study on vulnerability assessment for urban buried gas pipeline network based on SVM and ANN methods," *Process Saf. Environ. Prot.*, vol. 122, pp. 23–32, 2019.

[8] Z. Yan-Yan, Q. Bing, and L. Ting, "Integrating intra-and inter-document evidences for improving sentence sentiment classification," *Acta Autom. Sin.*, vol. 36, no. 10, pp. 1417–1425, 2010.

[9] H. Liu, D. Li, and Y. Li, "Poisonous label attack: black-box data poisoning attack with enhanced conditional DCGAN," *Neural Process. Lett.*, vol. 53, no. 6, pp. 4117–4142, 2021.

[10] D. Sharma, M. Sabharwal, V. Goyal, and M. Vij, "Sentiment analysis techniques for social media data: A review," in *First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019*, 2020, pp. 75–90.

[11] A. Jain, K. Bhatia , R. Sharma and S. Bhadola, "An emotion recognition framework through local binary patterns," Journal of Emerging Technologies and Innovative Research, Vol -6, Issue-5, May 2019, pp.675-684.



International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |