Educational Administration: Theory and Practice
2024, 30(11), 885 - 895
ISSN: 2148-2403
https://kuey.net/ **Research Article**

Techniques and Applications for Sentiment Analysis: A Perception

Deepak Kumar[1*], Dr. Bhavana Narain[2]

[1*]Research Scholar, MSIT, MATS University, Raipur, India, Email: deepakrsingh523@gmail.com
[2]Professor, MSIT, MATS University, Raipur, India, Email: narainbhawna@gmail.com

Citation: Deepak Kumar et al. (2024), Techniques and Applications for Sentiment Analysis: A Perception, *Educational Administration: Theory and Practice*, 30(11) 885 – 895
Doi: 10.53555/kuey.v30i11.8842

ARTICLE INFO

ABSTRACT

Sentiment Analysis (SA) is still being conducted in the field of text mining. SA is the computational handling of textual subjectivity, sentiments, and opinions. This survey article addresses a thorough summary of the most recent developments in this area. Numerous algorithms have been proposed recently. This review examines and briefly presents several SA applications and upgrades. These articles are grouped according on how they contribute to different SA approaches. The linked domains that drew attention to SA (transfer learning, emotion detection, and resource building) Recent researchers are discussed. This survey's primary goal is to provide a virtually complete picture of Brief descriptions of SA techniques and associated disciplines. This paper's primary contributions include the intricate classifications of numerous recent works and the example of the current research trend in sentiment analysis and related fields.

Keyword: Sentiment analysis, sentiment Classification, feature selection, emotion Detection, transfer learning.

I. Introduction

The computational examination of people's beliefs, attitudes, and feelings about an issue is known as sentiment analysis (SA) or opinion mining (OM). The entity may stand in for people, occasions, or subjects. Reviews are more likely to address these subjects. The terms SA and OM can be used interchangeably. They both convey a meaning. Nonetheless, several researchers claimed that SA and OM have rather different ideas [1]. Sentiment analysis finds and examines the sentiment represented in a text, whereas opinion mining gathers and examines people's opinions about a thing. Thus, as seen in Fig. 1, the goal of SA is to gather viewpoints, determine the emotions they convey, and then categorize their polarity. As seen in Fig.1, sentiment analysis can be thought of as a categorization procedure. Document-level, sentence-level, and aspect-level SA are the three primary classification levels in SA. Classifying an opinion document as expressing a favorable or negative opinion or sentiment is the goal of document-level SA. It views the entire document as a basic information unit (discussing a single subject). Classifying the sentiment conveyed in each sentence is the goal of sentence-level SA.
Determining whether the sentence is subjective or objective is the first stage. Sentence-level SA will ascertain whether a subjective sentence conveys favorable or unfavorable opinions. It has been noted by Wilson et al. [2] that sentiment expressions are not always subjective. But there isn't fundamental distinction between categorization at the document and sentence levels, as sentences are really brief documents [3]. Text classification at the sentence or document level does not offer the specific information required to form views on all features of the entity that are required for numerous applications, We must travel to the aspect level in order to get these facts. The goal of aspect-level SA is to categorize the sentiment in relation to the unique characteristics of entities. The initial stage is to determine the things and their characteristics. Opinion holders may express differing views on several facets of the same entity, such as this phrase "This phone's speech quality is poor, but the The battery lasts a long time. This poll addresses the initial two types of SA. One significant challenge in this discipline is the data sets employed in SA. Product reviews are the primary data sources. These For business owners, evaluations are crucial since they can take business choices based on the findings of user analysis views regarding their merchandise.

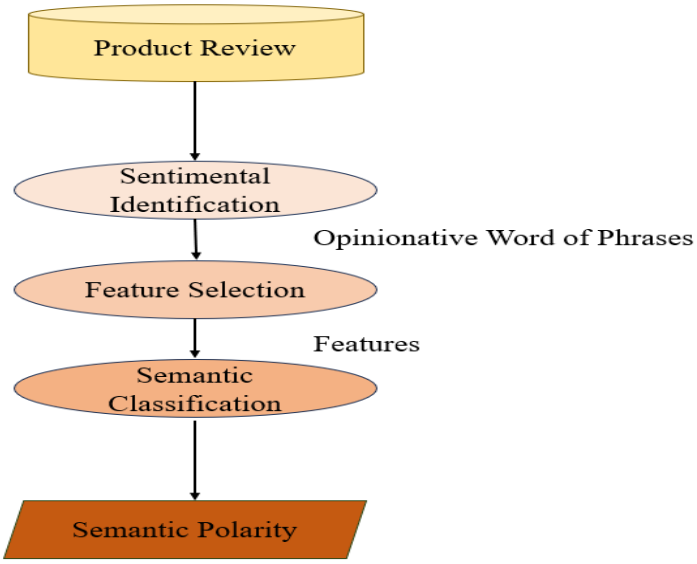

Fig. 1: Semantic Analysis on Product Review

The primary sources for the reviews are review websites. SA is used for more than just product reviews; it can additionally be used in news stories, stock markets [4,5], [6], or political discussions [4]. For instance, in political discussions, we could determine the public's perceptions of a certain election candidate or political organizations. Additionally, the election outcomes can be forecasted from positions in politics. Social media platforms and microblogging Websites are regarded as excellent information sources because People express and debate their thoughts on a certain subject without restriction. In the SA process, they serve as data sources as well.

On SA, there are numerous applications and improvements. algorithms that have been put forth in recent years. The purpose of this study is to examine these improvements in further detail and to List and classify a few articles that have been presented in this field. based on the several SA methods. The authors have compiled 54 articles that highlight significant recent advancements in the field of SA. These articles discuss a variety of a range of SA fields. All of them were released within the last several years. They are grouped based on the article's goal, which includes displaying the data and algorithms they employed. Fig. 1 shows that the writers have talked about the feature. Details of selection (FS) methods and their associated articles that make reference to a few original sources. The Feeling As illustrated in Fig. 2, classification (SC) procedures are explored in greater length, with examples from linked topics and originating allusions as well. This survey may prove beneficial for novice researchers in this field. field since it includes the most well-known SA methods and uses in a single study. This survey provides a unique improved classification to the several SA methods, which is absent from other surveys. It also covers recently developed linked topics. which have recently drawn the attention of scholars and the publications that go along with them in SA. Among these domains is the detection of emotions. (ED), Transfer Learning (TL), and Building Resources (BR).
The goal of emotion detection is to identify and evaluate feelings, whereas the sentences may contain explicit or implicit emotions. When it comes to cross-domain classification or transfer learning by examining information from one area and then applying the leads to the target domain. The goal of Building Resources is to create Lexica and corpora that include annotations for opinion expressions based on their polarity, and occasionally dictionaries. In this paper, the writers examine these areas in further detail. Each year, a large number of articles are given. in the fields of SA. The quantity of articles is growing as a result of years. Survey papers that provide an overview of the latest research trends and directions in SA are therefore necessary. The person who reads may locate a number of complex and thorough surveys, such as [1, 3, 8, 11]. The issue of SA has been covered in those surveys. from the perspective of applications rather than SA methodologies. Pang gave two lengthy and comprehensive surveys, and Liu [3], Lee [8]. They concentrated on SA's applications and difficulties. They discussed the methods for resolving every issue in SA. Feldman and Cambria and Schuller et al. [9] [10] and Martínez-Barco and Montoyo [11] have provided brief surveys that show the latest developments in South Africa. Tsytsarau as well as Palpanas [6] provided a survey that included the primary details about SA issues. They have provided illustrations for each topic. definition, issues, and progress, and grouped the articles that use graphs and tables.
The examination of the articles in this poll are comparable to those that were provided. by [5], although from a different angle and using a different classification among the articles. This survey makes a substantial contribution for a number of reasons. Initially, this study offers an advanced classification of numerous recent publications based on the methods utilized. Researchers who are familiar with this angle may find it useful. With specific methods to apply them in the SA domain and select the right method for a particular use case.

Secondly, Brief descriptions are provided for each of the different SA approaches. of the algorithms and the sources from which they were derived. This can give newcomers to the SA field a broad perspective. across the field. Third, the benchmark data sets that are accessible are examined and arranged based on their application in specific applications. Lastly, the survey is improved by talking about the linked domains to SA, such as identifying emotions, creating resources and learning transfer. The structure of this document is as follows: Section 2 contains the methodology for the survey and an overview of the articles. Part Three addresses the FS methods and the papers that are relevant to them, and the many SC approaches and the related articles are covered in Section 4. Section 5's relevant fields for SA and the articles that correspond to them are displayed. Section 6 provides the findings and debates, and ultimately the conclusion and Section 7 addresses research trends for the future.

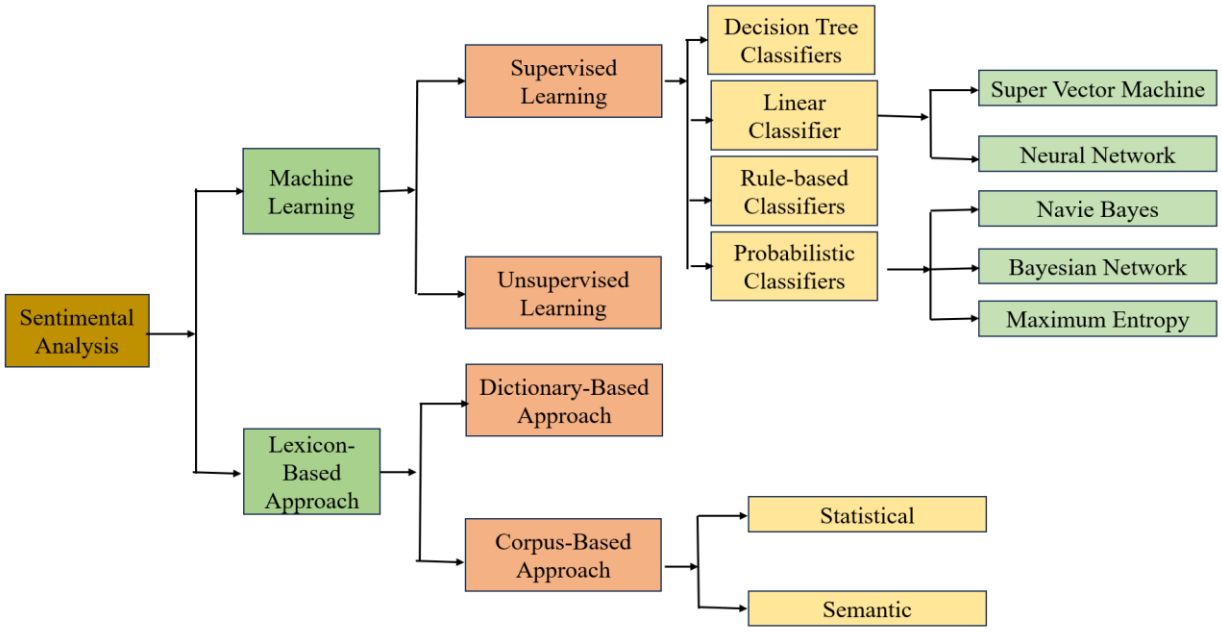

Fig. 2: Sentimental Classification Techniques

II. Methodology

Table 1 provides an overview of the 54 papers that were included in this survey. The papers referencing [4–7] and [12–8] are included in Table 1. The third column provides examples of the articles' goals. They fall into the following six categories: SA, ED, SC, FS, TL, and BR. The BR category can be categorized as dictionaries, corpora, or lexica. The articles that address the sentiment classification problem were classed as SC by the authors. SA refers to other articles that address a broad Sentiment Analysis issue. FS refers to the articles that contribute to the feature selection process. The writers then grouped the papers that represented SA-related domains such as Transfer Learning (TL), Building Resources (BR), and Emotion Detection (ED). Using Yes/No responses (Y or N), the fourth column indicates if the article is domain-oriented. Domain-oriented refers to the SA process's utilization of domain-specific data. As seen in Fig. 2, the fifth column lists the algorithms that were employed together with their classifications. Other algorithms beyond the SC approaches described in Section 4 are used in some works. This is relevant, for instance, to Steinberger's work [13]. Here, simply the name of the algorithm is written. The article's usage of SA techniques for general Analysis of Text (G) or its solution to the binary classification problem (Positive/Negative) are indicated in the sixth column. The extent of the data used to assess the article's algorithms is shown in the seventh column. Reviews, news articles, websites, microblogs, and other types of content could be included in the data. Since some articles do not provide that information, the eighth column details the benchmark data set or, if available, the reputable data source used. If the reader is interested in a certain range of data, this could be helpful. If the article analyzes any other languages except English, it is shown in the final column. The following is the survey's methodology: The well-known FS and SC algorithms, which represent some relevant fields to SA, are briefly explained. The articles' contributions to these algorithms are then shown, along with examples of how they are used to address unique SA challenges. Presenting a distinctive classification for these SA-related papers is the primary goal of this survey.

III. Choosing features for emotion categorization

The work of sentiment analysis is seen as a sentiment classification issue. Extracting and choosing text features is the first step in solving the SC problem. Currently available features include [21]: Word presence and frequency are characteristics that include the frequency counts of individual words or word n-grams. To show the relative importance of features, it either employs term frequency weights or binary weighting, which assigns a value of one if the word appears and zero otherwise [13]. Finding adjectives, which are significant markers of opinions, is part of the parts of speech (POS). Words and phrases that are frequently used to convey opinions include "good" or "bad," "like" or "hate." However, certain sentences convey viewpoints without the need of opinion words. Cost me a fortune, for instance. Negatives: the use of negative language can alter one's perspective, such as when "not good" is used interchangeably with "bad. "Techniques for feature selection the two categories of feature selection techniques are statistical techniques, which are more commonly employed and automatic, and lexicon-based techniques, which require human annotation. Typically, lexicon-based methods start with a limited number of "seed" words. To get a wider vocabulary, they then bootstrap this set using online resources or synonym identification. According to Whitelaw et al. [14], this turned out to have numerous challenges. Conversely, statistical methods are entirely automated.
The feature selection methods either handle the documents as a string that preserves the word sequence in the document or as a collection of words (Bag of Words, or BOWs). Due to its ease of use for the classification process, BOW is utilized more frequently. The removal of stop words and stemming—returning the word to its stem or root, such as "flies fi fly"—are the most popular feature selection steps. Three of the most popular statistical techniques in FS and the articles that discuss them are shown in the following subsections. Other techniques, such as the Gini index and information gain, are employed in FS [12]. Table summarized other techniques, such as the Gini index and information gain, are employed in FS [15].

3.1. Mutual Information Point-wise (PMI)

A formal method for modeling the mutual information between the features and the classes is offered by the mutual information measure. The information theory served as the basis for this metric [15]. The mutual information point-wise (PMI) Based on the degree of co-occurrence between the class I and the word w, Mi(w) between the two is defined. Pi F(w) represents the predicted co-occurrence of class I and word W based on their mutual independence, while F(w) p(w) represents the true co-occurrence. The mutual information is provided by the following formula and is defined as the ratio of these two values:

$$M_i(w) = log\ log\ \frac{F\ (w).p_i(w)}{F\ (w).p_i} = log\ log\ \left(\frac{p_i\ (w)}{p_i}\right) - - - - - - - - - - - - - - - (1)$$

When Mi(w) is greater than 0, there is a positive correlation between the word w and class I. When Mi(w) is smaller than zero, there is a negative correlation between the word w and class I. There are numerous uses for PMI, and it has been improved. Only the co-occurrence strength is taken into account by PMI. By creating a contextual entropy model to increase a collection of seed words derived from a limited corpus of stock market news items, Yu and Wu [4] have expanded the fundamental PMI. In order to find terms that are comparable to the seed words, their contextual entropy approach compares the contextual distributions of two words using an entropy metric. Following the expansion of the seed words, the sentiment of the news stories is categorized using both the expanded and seed terms. Their findings demonstrated that their approach may find more helpful emotion words and that their classification ability is enhanced by the intensity of those words. Their approach fared better than the (PMI)-based expansion techniques because it takes contextual distribution and co-occurrence strength into account, resulting in the acquisition of more useful emotion words and less noisy words.

3.2. Chi-Square (χ^2)

Let pi(w) be the conditional probability of class I for documents that contain w, let n be the total number of documents in the collection, Let F(w) represent the global proportion of documents that contain the word "w," and let Pi represent the global fraction of documents that contain the class "i." As a result, [16] defines the v2-statistic of the word between word w and class I:

$$\chi 2_i = \frac{n\ .F(w)^2\ .(p_i\ (w) - p_i)^2}{F(w).\big(1 - F(w)\big).p_i.(1 - p_i)} - - - - - - - - - - - - - - (2)$$

There are two distinct methods for calculating the correlation between terms and categories: v2 and PMI. Since v2 is a normalized number, it is superior than PMI since it allows for more comparability between terms in the same category [16]. Contextual advertising, as proposed by Fan and Chang [27], is one of the several applications of v2. To enhance online contextual advertising, they identified the bloggers' immediate personal interests. They worked on genuine blog entries and advertisements from epinions.com, wikipedia.com, and ebay.com. They employed v2 for FS and SVM (detailed in the following section) for classification. Their findings shown that their approach could successfully find advertisements that have a positive correlation with a blogger's individual interests. As part of their feature selection procedure for stock market data, Hagenau and

Liebmann [5] employed market input to create feedback features. They then applied them to Bi-Normal Separation (BNS) and v2.
They demonstrated how combining complicated feature types with a strong feature selection greatly improves classification accuracies. Their method lessens the issue of over-fitting when using a machine learning methodology and enables the selection of semantically meaningful characteristics. SVM was employed as a classifier. According to their findings, combining sophisticated feature extraction techniques with feedback-based feature selection improves sentiment analytics and classification accuracy. This is because, when using machine learning techniques to categorize text messages, their method minimizes the detrimental consequences of over-fitting and permits the reduction of the quantity of less-explanatory features, or noise.

3.3. Latent Semantic Indexing (LSI)

By selecting from the initial set of attributes, feature selection techniques aim to decrease the dimensionality of the data. As a function of the initial feature set, feature transformation techniques produce a reduced set of features. One of the well-known feature transformation techniques is LSI [16]. By combining the original word features linearly, the LSI technique converts the text space into a new axis system. To do this, Principal Component Analysis (PCA) techniques are employed [17]. It chooses the axis-system that preserves the most information regarding the changes in the values of the underlying attributes. LSI's primary drawback is that it is an unsupervised method that ignores the underlying class distribution. As a result, the characteristics identified by LSI may not always correspond to the paths that best distinguish the class-distribution of the underlying texts [18]. Other statistical techniques, such as Latent Dirichlet Allocation (LDA) and the Hidden Markov Model (HMM), could be applied in FS. Duric and Song [13] employed them to distinguish between the subjective expressions that characterize the entities in a review document in terms of polarity and the entities themselves. They suggested these new feature selection methods. Generative models known as LDA enable documents to be described by latent (unobserved) subjects. A topic model called HMM-LDA models themes and syntactic structures in a set of documents at the same time [18]. When utilizing just syntactic classes and minimizing overlaps with semantic words in their final feature sets, the feature selection strategies put forth by Duric and Song [23] produced competitive results for document polarity classification.
They employed the Maximum Entropy (ME) classifier while working on movie reviews; specifics are provided in the following section. Irony identification is a highly difficult feature extraction problem. Finding ironic reviews is the aim of this work. Reyes and Rosso were the ones who suggested this work [18]. In order to reflect a portion of the subjective information that underpins these reviews and attempts to characterize key aspects of irony, they set out to create a feature model. Six feature categories—n-grams, POS-grams, funny profiling, positive/negative profiling, emotional profiling, and pleasantness profiling—have been developed to describe verbal irony. They created a publicly accessible data collection including customer reviews, satirical pieces, and sarcastic reviews from news sources gathered from Amazon.com. They were shared because of an online viral impact, which is content that sets off a domino effect among users. For classification purposes, they employed NB, SVM, and DT (detailed in the following section). Their accuracy, precision, recall, and F-measure outcomes with the three classifiers are all satisfactory.

IV. Sentiment Classification Techniques

The three main categories of sentiment classification techniques are hybrid, lexicon-based, and machine learning approaches [19]. The Machine Learning Approach (ML) makes use of linguistic features and the well-known ML algorithms. A sentiment lexicon, which is a compilation of pre-compiled and known sentiment terms, is the foundation of the Lexicon-based Approach. Dictionary-based and corpus-based approaches, which employ statistical or semantic techniques to determine sentiment polarity, are the two categories into which it falls. The hybrid technique, which blends the two strategies, is widely used, with sentiment lexicons being essential to most strategies. As previously stated, Fig. 2 depicts the different strategies and the most widely used algorithms of SC.
The two main categories of machine learning-based text classification techniques are supervised and unsupervised learning techniques. Numerous labeled training materials are used in the supervised approaches. When it is challenging to locate these labeled training documents, unsupervised techniques are employed. Finding the opinion vocabulary that is utilized to analyze the text is the foundation of the lexicon-based method. This strategy consists of two methods. The dictionary-based method looks for synonyms and antonyms for opinion seed words after identifying existing ones. In order to find opinion words with context-specific orientations, the corpus-based approach starts with a seed list of opinion terms and then searches a huge corpus for further opinion words. Both statistical and semantic approaches could be used to accomplish this. The following subsections provide a brief description of the algorithms used in both techniques as well as relevant papers.

4.1. Machine Learning Approach

Utilizing syntactic and/or linguistic factors, the machine learning technique uses the well-known machine learning algorithms to tackle the SA as a standard text classification issue. Definition of the Text Classification Problem: Each training record in our collection, D = {X1, X2,..., Xn}, is assigned to a class. One of the class

labels in the classification model is connected to the characteristics in the underlying data. The model is then used to predict a class label for a specific instance of an unknown class. When an instance is given only one label, it presents a challenging categorization problem. When an instance is given a probabilistic label value, this is known as the soft classification issue.

4.1.1. Supervised Learning

The presence of labeled training documents is necessary for the supervised learning techniques to function. The literature contains a wide variety of supervised classifier types. We briefly describe some of the most popular classifiers in SA in the upcoming subsections.

4.1.2. Probabilistic Classifiers

Mixture models are used for classification by probabilistic classifiers. According to the mixture model, every class is a part of the mixture. Every component of the mixture is a generative model that gives the likelihood of sampling a specific phrase for that component. Another name for these classifiers is generative classifiers. In the next subsections, three of the most well-known probabilistic classifiers are covered.

4.1.2. Navi Bayes Classifier

The most straightforward and widely used classifier is the Naïve Bayes classifier. The posterior probability of a class is calculated using the Naïve Bayes classification model using the document's word distribution. The BOWs feature extraction used by the model disregards the word's location within the document. It forecasts the likelihood that a given feature set corresponds to a specific label using the Bayes Theorem.

$$P(label\backslash features) = \frac{P(label) * P(features\backslash label)}{P(features)} - - - - - - - - - (3)$$

The prior probability of a label, or the chance that a random feature established the label, is denoted by P(label). The prior probability that a particular feature set is being categorized as a label is P(features|label). The prior likelihood that a particular feature set is happened. Considering the naïve premise, which holds that ever Since characteristics are independent, the formula might be reformulated as follows:

$$P(label\backslash features) = \frac{P(label) * P(f1\backslash label) * \ldots\ldots\ldots * P\ (fn\backslash label)}{P(features)} - - - - - - - - - (4)$$

In order to address the issue of the tendency for the positive classification accuracy to appear up to roughly 10% higher than the negative classification accuracy, Kang and Yoo [36] suggested an improved NB classifier. When the accuracies of the two classes are expressed as an average value, this leads to a problem of diminishing the average accuracy. In comparison to NB and SVM, they demonstrated that applying this algorithm to restaurant reviews reduced the difference between the positive and negative accuracy. When compared to both NB and SVM, the accuracy is increased in both recall and precision.

4.1.3. Bayesian Network

The independence of the characteristics is the NB classifier's primary premise. The second extreme presumption is that every characteristic is completely reliant. Consequently, the Bayesian Network model is an acyclic directed graph with random nodes Conditional dependencies are represented by variables and edges. BN is regarded as a comprehensive model that captures the variables and their connections. A full joint probability distribution, then For a model, (JPD) over all variables is defined. Within the text mining, BN's computational complexity is highly costly; It is not commonly utilized as a result [20]. Hernández and Rodríguez [21] used BN to think about a real-world issue where the author's mindset is defined by three distinct target variables that are connected. The application of multi-dimensional Bayesian networks was suggested by them. classifiers. It combined the several target variables into a single categorization task to take use of the possible connections in between them. To take use of the semi-supervised domain, they expanded the multi-dimensional classification framework. Benefit from the vast amount of unlabeled data that is accessible in this situation. They demonstrated that their partially supervised the multifaceted method works better than the most popular SA methods, and their classifier is the most effective one. in a framework that is semi-supervised since it corresponds to the real foundational domain structure.

4.1.4. Maximum Entropy Classifier (ME)

Using encoding, the classifier (also called a conditional exponential classifier) transforms labeled feature sets into vectors. This Weights for each feature are then computed using the encoded vector, and they can be added to establish the most likely a feature set's label. The parameters of this classifier are set by a the joint features are combined using a set of X{weights}. They are produced by an X{encoding} from a feature-set. In Specifically, each C{(feature set, label)} pair is mapped by the encoding. to a vector. Next, the likelihood of each label is calculated. utilizing the equation that follows:

$$P(fs\backslash label) = \frac{dotprod\ (weights, encode\ (fs, label)}{sum\ (dotprod\ (weights, encode\ (fs, l) forlinlabels)} - - - - - (5)$$

Kaufmann [22] employed the ME classifier to identify parallel sentences across any pair of languages using a little quantity of training data. The other tools that were created to automatically extract parallel data from

non-parallel corpora either require a lot of training data or employ language-specific approaches. Their findings demonstrated that ME classifiers can generate insightful output for nearly any pair of languages. As a result, parallel corpora for numerous new languages may be possible.

4.1.5. Lenier Classifier

Vector A ¼ fa1..... ang is a vector of linear coefficients with the same dimensions as X ¼ fx1..... :xng is the normalized document word frequency. b is a scalar, and the feature space; the result of the linear p ¼ A:X þ b is the predictor, which is the result of the classifier that is linear. The separating hyperplane is the predictor p. among several classes. Numerous types of linear classifiers exist, including Support Vector Machines (SVM) are one of them [20,31]. This is a type of classifier that looks for good linear separators for several classes. Two of the most These well-known linear classifiers are covered in the following sections.

4.1.6. Support Vector Machine

Finding linear separators in the search space that are most effective at separating the various classes is the fundamental idea behind SVMs. Two classes (x and o) and three hyperplanes (A, B, and C) are shown in Fig. 3. Since hyperplane A represents the biggest margin of separation and every data point's normal distance is the largest, it offers the best separation between the classes. The sparse character of text, where few features are unimportant but are typically associated with one another and arranged into linearly separable categories, makes text data perfect for SVM classification [22]. By non-linearly mapping the data instances to an inner product space where the classes can be divided linearly with a hyperplane, SVM can create a nonlinear decision surface in the original feature space [23].

4.1.7. Neural Network

The basic unit of a neural network is a neuron, which is made up of several neurons. The vector overline Xi, which represents the word frequencies in the ith text, serves as a representation of the inputs to the neurons. Each neuron has a set of weights A that are used to calculate a function of its inputs, f(·). The neural network's linear function is pi ¼ A ~ Xi. The class label of Xi is supposed to be represented by yi in a binary classification issue, and the class label is obtained from the sign of the predicted function pi.

4.1.8. Decision Tree Classification

The training data space is hierarchically decomposed by the decision tree classifier, which divides the data according to an attribute value condition [26]. The existence or lack of one or more words is the condition or predicate. The data space is divided recursively until a minimum number of records that are used for categorization are present in the leaf nodes.

4.1.9. Rule Based Classification

A set of rules is used to model the data space in rule-based classifiers. The class label is on the right, and a condition on the feature set given in disjunctive normal form is on the left. The requirements are based on the term presence. Due to its lack of information in sparse data, term absence is rarely utilized.

4.2. Unsupervised Learning

Text classification's primary goal is to group documents into a predetermined number of groups. As previously shown, supervised learning uses a large number of labeled training documents to achieve that. Making these labeled training documents for text categorization might occasionally be challenging, but gathering the unlabeled documents is simple. These challenges are addressed by unsupervised learning techniques. In this subject, numerous research studies have been presented, including one by Ko and Seo [27]. They suggested breaking the texts up into sentences and classifying each one using a measure of sentence similarity and a list of keywords from each category.

4.2.1. Meta Classifier

The researchers frequently test their findings using one or more classifiers. The work that Lane and Clarke [26] have proposed is one of these articles. They offered a machine learning technique to address the issue of identifying documents with favorable or unfavorable ratings in media analysis. They encountered difficulties in achieving their objective due to the unequal distribution of positive and negative samples, modifications in the documents over time, and efficient training and assessment processes for the models. They worked with three sets of data that were produced by a media analysis firm. They categorized documents in two ways: determining whether favorability was present and determining whether favorability was positive or negative. To generate the data sets from the raw text, they employed five distinct feature categories. To identify the best classifier, they examined a number of them, including SVM, K-nearest neighbor, NB, BN, DT, a rule learner, and others. They demonstrated that while NB may suffer, performance can be enhanced by balancing the class distribution in training data.

4.2.2. Lexicon Based Approach

Many tasks involving sentiment categorization use opinion words. While negative opinion words are used to convey undesirable states, positive opinion words are used to convey desired states. Additionally, there are opinion idioms and expressions, which collectively make up opinion lexicon. The opinion word list can be compiled or gathered using three primary methods. The manual method takes a lot of time and is not often utilized. As a last precaution against errors resulting from automated methods, it is typically used in conjunction with the other two automated processes. The subsequent subsections introduce the two automated methods.

4.2.3. Dictionary Based Approach

outlined the dictionary-based approach's primary tactic. A limited collection of opinion words with established orientations is gathered by hand. This set is then expanded by looking for synonyms and antonyms in the renowned corpora WordNet [27] and thesaurus [28]. The following iteration begins once the newly discovered words are added to the seed list. When no new words are discovered, the iterative process comes to an end. Errors can be eliminated or fixed by manual examination once the procedure is finished.

4.2.4. Cropp Based Approach

The challenge of identifying opinion words with context-specific orientations is addressed by the corpus-based approach. Its techniques rely on a seed list of opinion words and grammatical patterns or patterns that occur together to identify additional words of opinion in a big corpus. Among these techniques were represented by McKeown and Hatzivas siloglou [29]. They began by using a list of seed opinion descriptors. in addition to a list of linguistic restrictions to find more the orientations of adjective opinion words. The limitations are for conjunctions such as AND, OR, BUT, EITHER-OR, etc. For instance, the conjunction AND indicates that conjoined adjectives typically share the same orientation. This concept is known as sentiment constancy, which in practice isn't always consistent. Additionally, there are adversative phrases like but, but which are marked as shifts in opinion. In order to determine whether two adjectives that are connected are the same or different. orientations, a vast corpus is used for learning. Next, Adjective linkages create a network, and clustering is done. on the graph to generate two-word sets: favorable and adverse.

4.2.5. Statistical Approach

The Semantic method uses distinct ideas to calculate word similarity and provides sentiment values directly. According to this theory, words that are semantically similar have similar sentiment values. For instance, WordNet offers many words semantic associations that are utilized to determine sentiment polarities. By iteratively adding synonyms and antonyms to the original set, WordNet may also be used to generate a list of emotion words. Then, the relative number of positive and negative synonyms for an unknown word could be used to determine the sentiment polarity of that word [26].

4.2.6. Lexicon Based and Natural Language Processing Techniques

Sometimes, lexicon-based approaches are used with Natural Language Processing (NLP) techniques to identify syntactical structure and aid in the discovery of semantic links [24]. Moreo and Romero [27] employed their suggested lexicon-based SA method after first using NLP techniques as a preprocessing step. An automatic focus recognition module and a sentiment analysis module that may evaluate user attitudes of subjects in news items using a taxonomy-lexicon created especially for news analysis make up their suggested system. In situations when informal language predominates, their findings were encouraging.

4.2.7. Discourse Information

Discourse has been more and more significant in SA in recent years. Both sentences and clauses inside a single sentence might contain discourse information. In [95,96], sentiment annotation at the discourse level was examined. Five categories of rhetorical relations—Contrast, Correction, Support, Result, and Continuation—as well as accompanying sentiment data for annotation have been employed by Asher et al. [25]. Opinion frames are a concept put forth by Somasundaran et al. [16]. Opinions and the connections between their objectives make up the elements of opinion frameworks [3]. In order to improve sentiment categorization, they have improved their work and looked into design decisions while modeling a discourse scheme [27].

V. Discussion and Result

In this part, we examine the patterns of researchers' use of different data, techniques, or SA tasks. According to their contributions in a variety of categories, the following graphs show the number of articles (which were shown in Table 1) throughout time. The number of articles that contribute to the six categories of SA tasks between years and the total number are shown in Fig. 4. This figure demonstrates that SA and SC continue to draw more researchers. It is evident that their contributions are the largest in the total count and about equal across years. Since ML techniques are new areas of search, academics have recently become interested in the related domains of ED, TL, and BR.

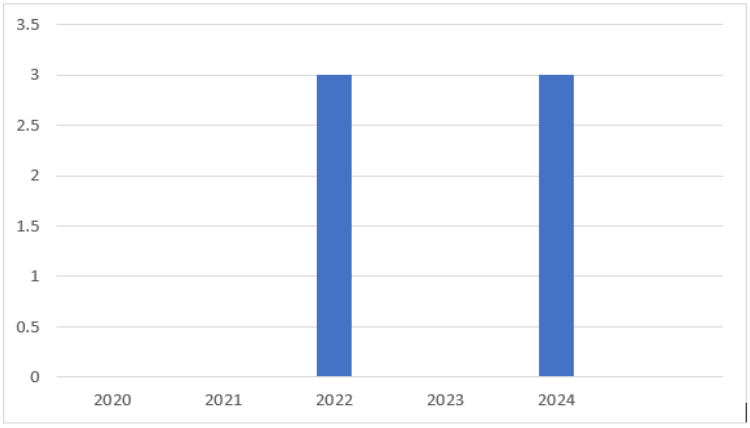

Fig. 3: Number of publications covering various sentiment analysis tasks over time

Because of their ease of use and capacity to leverage training data, which grants them domain flexibility, machine learning techniques are typically employed to address the SC problem. Due of their scalability, lexicon-based algorithms are widely utilized to address generic SA issues. Additionally, they are easy to use and computationally effective. In Fig. 5, the algorithms are displayed. As demonstrated, both the quantity and proportion of articles utilizing machine learning and Lexicon-based algorithms are evolving over time. Overall research over the past few years indicates that lexicon-based approaches are being used by researchers increasingly often. This is because, in spite of its enormous complexity, it can tackle a lot of SA jobs.
According to Tsytsarau and Palpanas [1], the majority of the work they presented used machine learning techniques, indicating that scholars have been moving toward broad text analysis in the last few years. Because hybrid approaches are more computationally complex, they are not yet widely used.

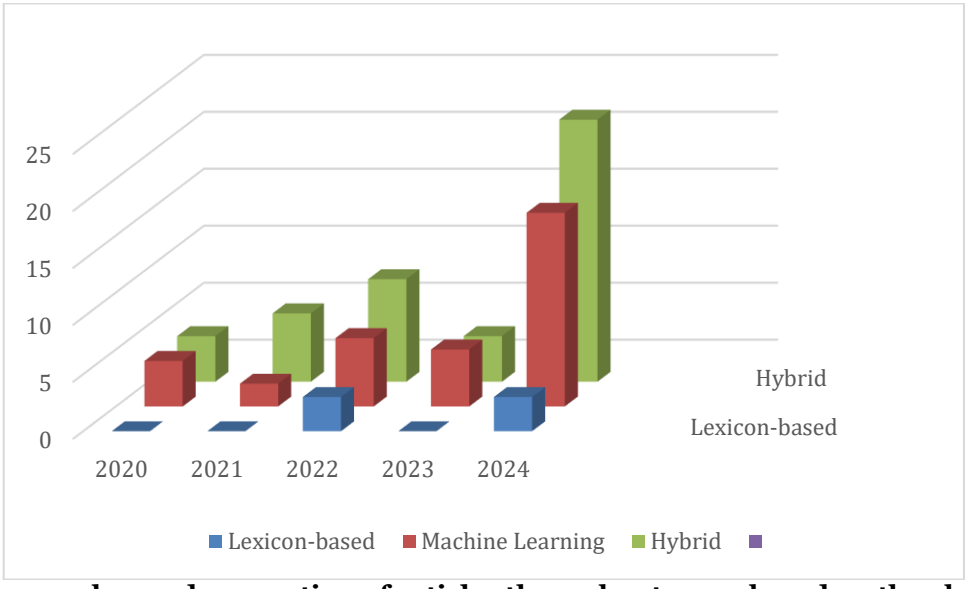

Fig. 4: The number and proportion of articles throughout years based on the algorithmic technique

VI. Conclusion and Future Work

An overview of the most current developments in SA algorithms and applications was provided in this survey study. We sorted and summarized 54 of the recently published and referenced publications. Numerous SA-related domains that employ SA techniques for a range of practical applications benefit from the contributions made by these papers. It is evident from examining these papers that there is still room for improvement in the SC and FS algorithms. The two most popular machine learning techniques for resolving SC problems are Naïve Bayes and Support Vector Machines. They are regarded as a reference model against which many suggested methods are evaluated. Since there are still few resources and studies on languages other than English, interest in these languages is developing in this sector. WordNet, which is available in languages other than English, is the most often used lexicon source. For many natural languages, building materials that are employed in SA tasks are still required. Recently, SA has made extensive use of information from news sources, blogs, forums, and microblogs. People's sentiments or opinions about a certain subject or product are often expressed through this media content. Further research is still required on the use of social networking and microblogging

websites as data sources. Some benchmark data sets are used for algorithm evaluation, particularly in reviews like IMDB. The user's preferences and the text's context are crucial factors in many apps. For this reason, more study on context-based SA is required. We can use training data that is relevant to the domain in question by employing TL approaches. Researchers have recently become interested in using NLP techniques to support the SA process, although there is still room for improvement.

Reference

[1] Tsytsarau Mikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Min Knowl Discov 2012;24:478–514.
[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP; 2005.
[3] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.
[4] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsuan-Shou. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl-Based Syst 2013;41:89–97.
[5] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Supp Syst; 2013.
[6] Tao Xu, Peng Qinke, Cheng Yinzhao. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowl- Based Syst 2012;35:279–89.
[7] Maks Isa, Vossen Piek. A lexicon model for deep sentiment analysis and opinion mining applications. Decis Support Syst 2012;53:680–8.
[8] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.
[9] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21.
[10] Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013;56:82–9.
[11] Montoyo Andre´ s, Martı´nez-Barco Patricio, Balahur Alexandra. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675–9.
[12] Qiu Guang, He Xiaofei, Zhang Feng, Shi Yuan, Bu Jiajun, Chen Chun. DASA: dissatisfaction-oriented advertising based on sentiment analysis. Expert Syst Appl 2010;37:6182–91.
[13] Lu Cheng-Yu, Lin Shian-Hua, Liu Jen-Chang, Cruz-Lara Samuel, Hong Jen-Shin. Automatic event-level textual emotion sensing using mutual action histogram between entities. Expert Syst Appl 2010;37:1643–53.
[14] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. Recognition of Affect, Judgment, and Appreciation in Text. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), Beijing; 2010. p. 806–14.
[15] Bai X. Predicting consumer sentiments from online text. Decis Support Syst 2011;50:732–42.
[16] Zhao Yan-Yan, Qin Bing, Liu Ting. Integrating intra- and interdocument evidences for improving sentence sentiment classification. Acta Automatica Sinica 2010;36(October'10).
[17] Yi Hu, Li Wenjie. Document sentiment classification by exploring description model of topical terms. Comput Speech Lang 2011;25:386–403.
[18] Cao Qing, Duan Wenjing, Gan Qiwei. Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach. Decis Support Syst 2011;50:511–21.
[19] He Yulan, Zhou Deyu. Self-training from labeled features for sentiment analysis. Inf Process Manage 2011;47:606–16.
[20] Tan Songbo, Wu Qiong. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Expert Syst Appl 2011:12094–100.
[21] Tan Songbo, Wang Yuefen. Weighted SCL model for adaptation of sentiment classification. Expert Syst Appl 2011;38:10524–31.
[22] Qiong Wu, Tan Songbo. A two-stage framework for crossdomain sentiment classification. Expert Syst Appl 2011;38:14269–75.
[23] Jiao Jian, Zhou Yanquan. Sentiment Polarity Analysis based multi-dictionary. In: Presented at the 2011 International Conference on Physics Science and Technology (ICPST'11); 2011.
[24] Lambov Dinko, Pais Sebastia˜ o, Dias Ga˜ el. Merged agreement lgorithms for domain independent sentiment analysis. In: Presented at the Pacific Association for, Computational Linguistics (PACLING'11); 2011.
[25] Xu Kaiquan, Liao Stephen Shaoyi, Li Jiexun, Song Yuxia. Mining comparative opinions from customer reviews for competitive intelligence. Decis Support Syst 2011;50:743–54.
[26] Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decis Support Syst 2011;50:755–68.
[27] Fan Teng-Kai, Chang Chia-Hui. Blogger-centric contextual advertising. Expert Syst Appl 2011;38:1777–88.

[28] Zhou L, Li B, Gao W, Wei Z, Wong K. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Presented at the 2001 conference on Empirical Methods in Natural Language Processing (EMNLP'11); 2011.
[29] Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F. Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM'11); 2011.
[30] Zirn C, Niepert M, Stuckenschmidt H, Strube M. Fine-grained sentiment analysis with structural features. In: Presented at the 5th International Joint Conference on Natural Language Processing (IJCNLP'11); 2011.
[31] Hu Nan, Bose Indranil, Koh Noi Sian, Liu Ling. "Manipulation of online reviews: an analysis of ratings, readability, and sentiments". Decis Support Syst 2012;52:674–84.
[32] Gupta Sunil Kumar, Phung Dinh, Adams Brett, Venkatesh Svetha. Regularized nonnegative shared subspace learning. Data Min Knowl Discov 2012;26:57–97.
[33] Duric Adnan, Song Fei. Feature selection for sentiment analysis based on content and syntax models. Decis Support Syst 2012;53:704–11.

Author Detail

Mr. Deepak Kumar
Asst. Professor, Computer Science and Application, Shri Shankaracharya Professional University, Bhilai, Chhattisgarh, India

Prof. (Dr.) Bhavana Narain
Professor, MATS School of Information Technology
MATS University, Raipur, Chhattisgarh, India

Prof. (Dr.) Bhavana Narain, Professor, MSIT, MATS University Raipur. She has **23** years of work experience in the academic area and **04** years includes industrial experience. She has completed her Ph. D. in Computer Science and Application. Her research domain is Digital Image Processing, Wireless Networking and Cyber Security. She has undergone two minor projects as Co-PI and one in house project as PI. She has worked as a State student coordinator of Region 4 Chhattisgarh state in the computer society of India, from 2014 to 2020. Currently, she has been elected as secretary of the CSI Raipur chapter from 2021 to 2024.
She has published **fourteen books** in international and national publication. She has published more than **82 papers** in national and international journals. She has presented papers in national and international conferences, **seven patents** and life membership of **six societies**. She has been awarded by eight national and international bodies. She has delivered lectures as a keynote speaker and has been working as session chair in more than **four** organizations. Editor, reviewer and expert speaker in national and international platform. She is member of the University Board of Studies.
Nine PhD scholars and **Four** M.Phil. CS dissertations have been awarded under her supervision. Presently, **Three** PhD thesis has been submitted under her supervision and guiding **six** PhD research scholars.