IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

REVIEW ON TECHNIQUES IN SENTIMENT ANALYSIS

[1]Sibi Mathew, [2]Abin P Mathew

[1,2]M.tech Scholar
[1,2]Department of CSE
[1,2]TKM College of Engineering,Kollam,Kerala,India,691005

Abstract : Sentiment analysis (SA), is the process of extracting emotions and opinions from the review of products and situations to understand the feelings from these text inputs. The rapid development of Internet-based applications such as websites, social networks, and blogs has allowed people to generate vast amounts of opinions and reviews about products, services, and everyday activities. Sentiment analysis is used by businesses as a great marketing strategy, governments to obtain a general public opinion, and researchers can extract and analyze public opinion and use it to gain business insights and make better decisions about a situation. The ongoing research focuses on expanding the scope of sentimental analysis through the use of a large chunk of data available on social media. In this paper, we will be looking in detail at the different techniques in sentimental analysis, its possible challenges, and current research in this field. This review intends to provide an overview of techniques used in sentimental analysis.

IndexTerms **- Natural Language Processing, sentiments, social media analysis, sentiments, techniques**

I.INTRODUCTION

Natural Language Processing (NLP) technique called sen timent analysis is used to glean feelings and viewpoints from the text. Natural language processing uses computational techniques to learn, understand, and generate human language content. Early computational approaches to language research focused on automating the analysis of the linguistic structure of language and developing underlying technologies such as machine translation, speech recognition, and speech synthesis.[5] Furthermore, new approaches to sentiment analysis are starting to include information from the text and other domains, such as visual data. It is possible to think of the process of identifying whether a text displays a positive or negative attitude as a classification issue. However, despite the fact that sentiment analysis can seem simple, it really involves analyzing a variety of NLP sub-tasks, including sarcasm and subjectivity detection. Additionally, there may be various grammatical flaws, slang terms, and acronyms, and the content is not always formatted like in books or newspapers. Today, sentiment assessment is usually accepted by corporations, ministries, and businesses as well as academics. The Internet has become the most essential and accessible source of information as a result of its growing popularity. Millions of individuals share their opinions and feelings via discussion groups, bloggers, community forums, social media, and other internet services. Since these opinions and feelings are so crucial to how we live our lives, it is important to analyze this reader data in order to automatically track public sentiment and support decision-making. For instance, exit polls have been predicted using tweets[1]. Global data show that there are more than 4 billion users in such social media, this can generate a variety of information through sentimental analysis. This availability of data has motivated researchers to work in this field of research. Sentiment analysis has become one of the fastest-growing and most active research fields, as the number of articles focusing on sentiment analysis and opinion mining has exploded in recent years [2]. This wide variety of research points out the importance of this field This paper gives a more comprehensive examination of the sentiment analysis as it discusses this area from several perspectives due to the fact that it contains numerous research components relating to emotion including difficulties, applications, tools, and strategies. As a result, researchers and beginners can benefit greatly from this as they get a tremendous amount of information regarding this subject in one paper. In order to produce the best outcome from this method of analysis a good understanding of the techniques is necessary

II.COMPUTATIONAL FLOW FOR ANALYSIS OF SENTIMENTS

In our review, we looked at a general computational flow that can be used to address the problem of sentiment analysis. The probable steps that can be used in the sentimental analysis can be briefly explained as follows:
(i) Text or Data collection step: Collecting text information is the initial stage in sentiment analysis, and there are many resources and tools accessible for this purpose. Text data is frequently produced or obtained for research purposes, by a third party, or via

web crawling and scraping. Improving textual information with other forms of data (such as telephony data, geo-location data, and video data) in order to do sentiment analysis can produce intriguing results.
(ii) Pre-processing: This stage aims to minimize the resolution of the input data because many words are unnecessary and must be eliminated because they have no bearing on the text polarity.
(iii) Feature Extraction: The goal of this activity is to extract pertinent information (such as words expressing a sentiment) that highlights important details in the text.
(iv) Classification step: The previously acquired features are used in a classification model that classifies them based on the number of distinct emotions. The objective is to improve the performance of sentiment analysis and to identify solutions to problems in this field.[10]

III.SENTIMENT ANALYSIS TECHNIQUES

Sentiment analysis is a thriving and active research field that can be applied to numerous domains. Therefore, researchers continually propose, evaluate, and compare various approaches. The objective is to improve the performance of sentiment analysis and to identify solutions to problems in this field. The improvement of performance comes through analysis of the different techniques used in sentimental analysis. This review is intended to provide an overview of the most common sentiment analysis techniques[6]. The techniques can be broadly classified into Machine Learning approaches[7], Lexicon-Based approaches[8], and Hybrid approaches [9]. Out of these machine learning becomes one of the most common and popular methods while the lexical approach uses a method similar to a bag of words and the hybrid method is a combination of the other two methods.

3.1 Machine learning approaches

Machine learning methods are used to categorize sentiment polarity based on the train and test datasets (e.g., negative, positive, and neutral). For this, four different strategies are used: reinforcement learning, semi-supervised learning, unsupervised learning, and supervised learning. The supervised strategy is used when a classification includes a predetermined set of classes; however, the unsupervised method may be used when it is challenging to identify this set due to the lack of labeled data. On the other hand, the semi-supervised approach can be used with unlabeled datasets that include some tagged examples. Trial-and-error techniques are used by reinforcement learning algorithms to help the agent interact with its environment and maximize cumulative rewards. To be effective, these methods usually need sizable training datasets. The performance of a classifier trained on a certain dataset, [11]however, is inferior to that of a classifier trained on a dissimilar domain.

3.1.1 Supervised learning

Labeled training materials are required for supervised techniques, and the classes are often used as the labels (e.g., positive, neutral, negative). There are four different categories of supervised classification techniques: decision trees, rule-based, probabilistic, and linear methods. The most popular supervised classification algorithms for sentiment analysis are briefly explained and contrasted in the sections that follow.

3.1.1.1 Linear approach

A statistical technique for classifying sentiment depending on linear or hyper-plane decision boundaries is known as a linear approach. The linear classifier $p = A.X + b$, which determines the text's classification upon the characteristics given, carries out this classification. The vector of linear coefficients (weights) and document frequency of the words are denoted by the letters A and X, respectively. In addition to bias b, the prediction is the dot product of A and X.

3.1.1.1.1 Support Vector Machine (SVM): is a non-probabilistic classifier that can manage both discrete and continuous variables and can divide data linearly or non-linearly. It has a strong theoretical base and, in many cases, outperforms the majority of existing classification algorithms in terms of accuracy. Finding the ideal hyperplane for class labels is the SVM classifier's main goal. An effective separation indicates that the hyperplane has the biggest margin to the closest training point from either class because a higher margin lowers the classifier's generalization error.

3.1.1.1.2 Artificial Neural Network (ANN): Recently, ANN has become well-known as a key classification technique. Its foundation is the idea that features can be retrieved through linear combinations of data input and then used to represent the result as a non - linear function of those features. Each of the three layers of a neural network's standard architecture—input, output, and hidden—contains a significant number of structured neurons. Neuronal connections within either layer serve to connect two subsequent levels.

3.1.1.2 Probabilistic approach

A probabilistic classifier, which is often based on Bayes' theorem, forecasts a probabilistic model over a set of classes. A probabilistic classifier predicts a probabilistic distribution over a set of classes as opposed to the linear technique, which outputs the most likely classification of a given data (either positive or negative). Each category is a component of the mixture as the classifier does classification using mixture models. In comparison to other methods, probabilistic classifiers are easy to use, computationally efficient, and don't need a lot of training data. Nevertheless, if the data do not (almost) match the distribution assumptions, classification performance may deteriorate.

3.1.1.2.1 Naive Bayes (NB): NB is a simplistic classifier that is among the most often used algorithms for classifying texts. The Bayes Theorem and BoW feature extraction underpin the model. As a result, it is irrelevant where a word appears within a document; instead, its presence stands alone from that of other words. In order to maximize P(c|d), naive Bayes applies the Bayes' rule and assigns document d to category c.

3.1.1.2.1.1 Bayesian Network (BN): Due to its flexible structure, it is straightforward to add new variables. It is a complete framework for representing the correlation between two of parameters using the joint distribution of probabilities. Bayesian networks were used in the text categorization field to find connections among a large number of words. The Naive Bayes, Support Vector Machine and C4.5 algorithms can be used to classify sentiment in a system that uses the Bayesian Network.

3.1.1.2.1.2 Maximum Entropy (ME): is a conditional exponential classifier also known as a Maxent classifier, that does not make any assumptions about the correlations between features. Naive Bayes and Maxent can be used to extract features from POS tags, which depend upon the Naive Bayes and Maxent classifiers, in order to increase the system's entropy. Given a document d, it calculates the conditional distribution of the target class c given a document d by using the exponential form. When used on unigrams and bigrams, the Maxent classifier exhibits astounding accuracy.

3.1.1.3 Rule-based approach

Any classification system that predicts class using IF-THEN rules is referred to as rule-based classification. In order to accomplish sentiment classification, this method's classifiers utilize a set of rules. This is how a rule might be expressed: LHS — > RHS. The LHS is the precursor of the rule or a set of constraints on the feature set written in the Disjunctive Normal Form, and the RHS denotes a conclusion of the rule if LHS is satisfied. Rule- based classifiers perform similarly to decision trees in terms of classification speed for new cases. The rule- based approach's ability to avoid overfitting is another advantage. However, when there are too many rules, they become laborious and difficult to interpret. Additionally, it struggles with noisy data. [12]

3.1.1.4 Decision tree approach

This technique decomposes the training data space hierarchically based on the attribute value in order to classify input data into a finite number of predefined classes. The attribute value condition is the presence or absence of one or more words. This based tree approach is a structure resembling a flowchart, where each internal node represents a test on an attribute, each branch represents the result of the test, and leaf nodes represent child nodes or class distributions. In addition to being simple to comprehend and interpret, decision tree classifiers can handle noisy data. However, they are unstable and susceptible to overfitting. The decision tree method performs exceptionally well with large datasets but is not recommended for small datasets.

3.1.1.5 Ensemble learning approach

The primary objective of this strategy is to combine multiple individual classifiers in order to produce a classifier that outperforms them all. When humans want to make an important decision, they apply this principle and take into account multiple opinions. This technique utilises all classifiers to make a more informed decision. In general, the final decision is reached through the application of a set of rules, such as the Majority Vote method. Due to the collaboration of the classifiers, an ensemble system tends to have better generalisation and accuracy, but the main drawback of this technique is that it requires more computation and training time than a single algorithm. Therefore, it is prudent to carefully select algorithms (e.g., quick algorithms such as decision trees).[13]

Boosting technique is an additional intriguing sequential method belonging to the family of ensemble methods. It involves training a series of weak classifiers to improve the performance of prediction. Each new classifier is only trained on samples that were incorrectly classified by its predecessors. The benefit of such a method is that the final classifier (a team of classifiers) is taught to make accurate predictions for all data types. Different boosting models, such as Adaptive Boosting (AdaBoost), Gradient Boosting Machine (GBM), and Boosted SVM, have been proposed.

3.1.2 Unsupervised learning

The majority of existing approaches for sentiment analysis rely on supervised learning models trained from labelled corpora in which each document is labelled prior to training. However, it can be challenging to collect and create labelled datasets, particularly for textual data, which is typically unstructured. In contrast, it is simpler to collect unlabeled datasets and classify them using unsupervised learning techniques. These techniques utilise statistical properties of documents, such as word co-occurrence, NLP processes, and existing lexicons containing emotional (or polarised) words. In machine learning, however, unsupervised approaches in the field of sentiment analysis typically employ clustering, which can classify data into different categories without specifying which sentiment each category represents. In other words, the clustering method divides data into groups (clusters) in which the data within a cluster are more similar than the data within different clusters. Cluster analysis techniques can be classified as Hierarchical or Partitioned.

3.1.2.1 Hierarchical methods

Hierarchical methods create a hierarchical decomposition of a dataset represented by nested clusters (groups with subgroups) organised in a tree structure. Hierarchical techniques can be divided into two primary strategies: Agglomerative clustering and Divisive clustering. This divisive clustering approach is known as the top-down method. This method begins with a single cluster that groups all the data and then assigns those data to sub-clusters based on their similarity through a recursive process.

Agglomerative clustering (also known as the bottom-up method) considers that each piece of data begins in its own cluster and then merges clusters containing similar data until only one or a small number of clusters remain.

3.1.2.2 Partition methods

The objective of partitioning methods is to partition data into a collection of non-overlapping clusters in which each element is assigned to a single cluster. This partitioning is determined by a similarity criterion that is typically the Euclidean distance between elements.
The k-means algorithm and its variants are the most popular algorithm for partitioning. The K-means algorithm begins with a predetermined number of initial cluster centroids and assigns iteratively the data objects in the dataset to cluster centroids based on their similarity to the cluster centroids. When a convergence criterion is met, the procedure comes to a close. The criterion can be a fixed number of iterations, or the result cannot change after a predetermined number of iterations.

3.1.3 Semi-supervised learning

When obtaining labelled data is difficult, semi-supervised learning (SSL) techniques are also employed. However, unlike unsupervised approaches, SSL employs a small set of initial labelled training data to guide the feature learning procedure. Therefore, it falls between supervised and unsupervised methods. SSL approaches make extensive use of large quantities of inexpensive unlabeled data, save a great deal of time and effort, and gain a classifier with strong generalizability in addition to more labelled data. SSL-based sentiment analysis techniques can be categorised as generative, co-training, self-training, graph-based, and multi-view learning.

3.1.3.1 Generative approach

This method assumes that data in different categories follow different distributions and that the parameters of each distribution can be estimated if at least one labelled data per category is available. After training this model for each class, a generative model predicts the label (class) of a test input using the Bayes rule.

3.1.3.2 Co-training approach

This method assumes that data can be represented using two independent views, each of which contains information about each data. In co-training, two distinct classifiers will be taught to instruct each other based on the information they share during training. Each classifier was trained on a different set of features corresponding to the two data views. At each iteration of the training process, co-training updates the dataset by adding the most confidently classified instances from each classifier to the labelled data. When all unlabeled data have been used or a predetermined number of iterations have been reached, the process terminates. This approach was used to propose a dual- view co-training approach that addressed the negation problem and improved bootstrapping efficiency for semi-supervised sentiment classification.

3.1.3.3 Self-training approach

This method is frequently employed for semi-supervised learning. The self-training process consists of two phases. Initially, a small amount of labelled data is utilized to train the classifier. In the second step, the trained classifier is used to classify unlabeled data in order to add the most confident samples as new labelled data to the original training set. The final step will be iteratively repeated with the newly labelled data. Using the test data, the resulting model is then evaluated. This method has been widely implemented in the field of sentiment analysis.

3.1.3.4 Graph-based approach

Graph-based approach, a graph architecture is used to represent the data. In a graph, vertices represent instances (e.g., sentences), whereas edges describethe similarity between instances. Strongly connected instances are typically members of the same class. Due to the extensive use of this methodology, its efficacy has been demonstrated in numerous NLP tasks, including sentiment analysis.

3.1.3.5 Multi-view learning approach

This approach takes multiple perspectives into account to solve a problem, and the overall performance is determined by their congruence. Each classifier will be trained on a single view and then used to label the unlabeled samples that will be added to the training set if they are reliably classified. Typically, this method is applied to problems with multiple distinct feature sets.

3.1.4 Reinforcement learning

Reinforcement learning (RL) is a machine learning technique in which an agent is rewarded based on the evaluation of its previous action in the subsequent time step. The algorithms of RL employ trial-and-error mechanisms to assist the agent in interacting with its environment to maximize cumulative rewards. Reinforcement learning has been used to solve a variety of problems, such as robot control, but its primary application has been in video games. Despite its capacity to handle complex tasks, particularly with the incorporation of Neural Networks, this technique is rarely used to solve problems involving sentiment analysis.[13] This method's primary advantage is its similarity to the human learning process, which is highly desired in the field

of sentiment analysis. Reinforcement learning employs learned historical experiences to correct errors made during training, and as a result, it makes better decisions and approaches perfection. On the other hand, the reinforcement learning model's conception can be laborious. In addition, reinforcement learning requires a large amount of data and is computationally costly.

3.1.5 Deep learning

Applying ANNs-based deep learning (DL) to sentiment analysis has recently become extremely popular. DL is an emerging field of machine learning that offers supervised and unsupervised methods for learning feature representation. Deep learning refers to neural networks with multiple perceptron layers inspired by the human brain. With this architecture, it is possible to train more complex models on a much larger dataset and produce state-of-the-art results in many application domains, including computer vision, speech recognition, and natural language processing.[14]

CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), and DBN (Deep Belief Networks) are a few of the neural network models included in deep learning. These models can learn sophisticated features from the dataset on their own. They are, however, computationally intensive and complex. The subsequent subsections, however, provide a brief description and summary of the most prevalent deep-learning models used for sentiment analysis.

3.1.5.1 Deep neural networks (DNN)

This model is an Artificial Neural Network (ANN) with hidden layers between the input and output layers. The input layer contains input data, the hidden layers contain processing nodes referred to as neurons, and the output layer contains one or more neurons used to produce network outputs. It employs sophisticated mathematical modeling and the learning ability of ANN to determine whether a mapping from input to output should be linear or nonlinear. The flow processes of ANNs and DNNs can be classified as feed-forward and feedback. Since feed-forward ANNs are simple networks, they are suitable for sentiment classification.

3.1.5.2 Convolutional neural networks (CNN)

This architecture is a subtype of feedforward neural network that was originally employed in the field of computer vision, but it has recently achieved success in other fields such as recommender systems and natural language processing. A CNN's layers include an input layer, an output layer, and a hidden layer composed of multiple convolutional layers, pooling layers, normalisation layers, and fully connected layers. Convolutional layers filter the inputs (such as word embedding in text sentiment classification) to extract features, whereas pooling layers reduce the resolution of features to make feature detection insensitive to noise and small changes. The normalisation layer normalises the output of a preceding layer in order to improve convergence during training, and the fully connected layers are used to perform the classification task. CNNs have recently gained widespread recognition in the field of sentiment analysis**.**

3.1.5.3 Recurrent neural networks (RNN)

This model processes a sequence of inputs using memory cells. RNNs are widely used in NLP tasks such as sentiment analysis due to their capacity to capture and recall information about long sequences. The output of RNNs is dependent on all previous computations. For instance, to predict the next word in a sentence, the model uses the states of all previous words and their relationship. A new type of RNN known as Long-Short Term Memory (LSTM) was introduced to combat the issue of vanishing gradient, which is one of the primary issues with standard RNNs. LSTM is gaining popularityin a variety of fields.

3.1.5.4 Additional neural networks

There are a number of other types of deep neural networks used for sentiment analysis, but not to the same extent as the three models listed above. Among them is the Recursive Neural Network (RecNN) model, which was employed to introduce a novel Recursive Neural Deep Model. This model outperformed Naive Bayes, Maximum Entropy, and SVM in classifying the binary sentiment of Chinese social data. Other unsupervised deep neural networks, such as Autoencoders and their variants, are also used in sentiment analysis, but it is difficult to employ them directly for this task. In general, the encoder layer is used as a feature extractor for classifiers.

3.2 Lexicon-based approach

The lexicon-Based (also known as knowledge-based) approach is one of the two primary approaches used for sentiment analysis and requires a lexical resource named opinion lexicon (a predefined list of words) that associates words with their semantic orientation as negative or positive words using scores. A score can be a simple polarity value, such as +1, 1, or 0 for positive, negative, or neutral words, or it can be a value that reflects the sentiment's strength or intensity. A document's final orientation is determined by calculating the semantic orientation values of its constituent words. A document is tokenized into single words or micro phrases, and each element is then assigned sentiment values from the lexicon. Formulas or algorithms (e.g., sum and average) can be utilized to determine the overall sentiment of a given document.[8] The lexicon-based approach to sentiment analysis at the sentence and feature levels is highly applicable. It does not require training data and is therefore considered an unsupervised method. As words can have multiple meanings and senses, a word that is positive in one domain may not be in another. For instance, given the word "small" and the sentences "The TV screen is too small" and "This camera is very small," the word "small" in the first sentence is negative because most people prefer wide screens, whereas in the second sentence, it is positive because a small camera will be easy to carry. Creating a domain-specific sentiment lexicon or employing a lexicon

adaptation strategy can circumvent this issue. If a large training dataset is provided, the performance of the lexicon-based approach decreases in comparison to the machine learning approach, which is another drawback of this method. Below are the three primary techniques for creating and annotating sentiment lexicons.

3.2.1 Manual Approach

The manual method necessitates human intervention for lexicon annotation. The creation of sentiment lexicons consists of two phases: the generation of a list of words that convey a particular emotion and the assignment of emotion labels to these words. This is typically a very laborious, expensive, and time-consuming process, but it can yield a consistent and reliable lexicon. To expedite this procedure, an automated method can be utilized. In this instance, a manual approach is utilized as a benchmarking procedure or to reduce the number of errors. Numerous lexicons have been manually compiled, which include MPQA Subjectivity Lexicon and Semantic Orientation CALculator (SO-CAL), which are based on manual lists of negators and intensifiers. Crowdsourcing and gamification may also be employed by researchers. Crowdsourcing is the practice of utilizing the Internet to recruit a group for a common purpose. In contrast, gamification is the application of game mechanics to non-game problems.

3.2.2 Dictionary based Approach

This approach is based on the premise that synonymous words have identical sentiment polarities, while antonyms have opposite polarities. Using well-known dictionaries such as WordNet or thesauri the sentiment lexicons in this method are compiled. Manually collecting a list of initial seed words with known orientations is the first step. The following step involves searching for the words' synonyms and antonyms in additional lexical resources. The newly discovered words are iteratively added to the previous list until no additional words are discovered. Later, a manual examination can be performed to identify and correct errors. Such a well-known lexicon is the SentiWordNet 3.0 was developed, by annotating all WordNet 3 synsets automatically.

The primary issue with dictionary-based and Lexicon-based approaches is their inability to locate sentiment words with domain-specific orientation; consequently, they are unsuitable for context- and domain-specific classification. Moreover, compiling dependency rules is difficult and laborious, but on the other hand, this technique is computationally inexpensive as long as there is no dataset training and represents a good strategy for easily and rapidly building a lexicon containing a large number of sentiment words and their orientation.

3.2.3 Corpus-based method

Unlike dictionary-based approaches, corpus-based approaches begin with a list of seed sentiment words whose orientation is already known and then exploit syntactic or co-occurrence patterns to find new sentiment words with their orientation in a large corpus. Utilizing linguistic constraints or conventions on connectives, additional sentiment words are identified (e.g., AND, OR, BUT). For instance, a pair of adjectives joined by a conjunction (such as "simple AND easy") typically have the same orientation. In addition to this concept, known as sentiment consistency, which is not always consistent in practice, a set of rules can be formulated for these connectives. Upon completion of this procedure, a number of techniques, such as clustering, can be used to create sets of sentiment words (e.g., positive and negative words) . The primary benefit of the corpus-based approach is its simplicity; however, it requires a large dataset to detect the polarity of words and, consequently, the sentiment of the given text . The corpus-based approach is usually divided into statistical and semantic approaches as described in the following subsections.

3.2.3.1 Methodological statistics

This method determines the emotional orientation of a word based on the statistical concept. This approach is based on the premise that similar sentiment words typically have the same meaning if they frequently appear in the same context. Therefore, the unknown polarity of a word is determined based on its co-occurrence frequency with other words that appeared in the same context. The co-occurrence frequency is computed using Turney's method for calculating mutual information. They utilize shared knowledge to associate terms with their respective PoS tags in the lexicon. This methodology has been utilised to generate sentiment lexicons and conduct sentiment analysis.

3.2.3.2 Semantic strategy.

This technique (also referred to as an ontology-based approach) employs different rules to measure the similarity between words and assigns the same sentiment value directly to semantically similar words . This method searches sentiment dictionaries for synonyms, antonyms, and words with similar meanings to extend a lexicon and perform sentiment analysis.

3.3 Hybrid approach

The hybrid method combines lexicon and machine learning methods. It combines the speed of lexical analysis with the adaptability of machine learning techniques to handle ambiguity and incorporate the context of sentiment words. The primary purpose of the hybrid approach is to inherit the high precision of machine learning and the consistency of the lexicon-based approach. The hybrid strategy combines techniques from the two preceding strategies to overcome their limitations and capitalize on their benefits. The lexicon approach scores are therefore used as input features for the sentiment classifier. Consequently, sentiment lexica play a crucial role in the hybrid approach, which is typically known for its superior performance.[16], [17] Only a few models employ the hybrid method for sentiment analysis. A well-known hybrid model uses machine learning classifiers with dictionaries and the HARN algorithm, which is proposed in a lexicon-based approach to document classification. In this

model, they first classify the reviews of each domain using two machine learning classifiers, namely Naive Bayes and SVM and then identified the polarity at the document level using HARN's algorithm. The hybrid method was approximately 80–85 percent more accurate than HARN's algorithm alone. Deep Learning can also be used in conjunction with lexicons for sentiment analysis. Such a model was introduced by creating lexicon embeddings by combining word scores from numerous lexicon sources. Three methods are used to incorporate these embeddings into a CNN model: naive concatenation, multichannel convolution, and separate convolution. This integration showed that lexicon integration can enhance the CNN model's precision, stability, and efficacy. By combining machine learning and lexicon- based techniques, a new hybrid method was introduced for identifying the sentiment polarities of tweets. SVM, Logistic Regression, and Recurrent Neural Network (RNN) classifiers were used for classification and they created a words lexicon, emoticon lexicon, idioms lexicon, and some essential lexica.

3.4 Other approaches

3.4.1 Aspect-based approach

Aspect-based sentiment analysis is a task for fine-grained sentiment analysis that aims to predict the sentiment polarities of the given aspects or target terms in texts (e.g., product or service). Aspects can be attributes, qualities, or characteristics of the target. Aspect-based sentiment classification consists of two steps: aspect extraction and classification of sentiment. The first stage extracts and groups synonyms of aspects used to refer to the same entity, and the second stage determines the sentiment of each extracted aspect.

3.4.2 Transfer learning

Transfer learning is the method that employs similarity of data, data distribution, model task, etc., to apply previously acquired knowledge from one domain to another. This method has emerged as a new technique for machine learning, and it is extremely useful, especially for saving time because it eliminates the need to train an algorithm from scratch. Transfer learning is typically used in sentiment analysis to transfer the acquired ability to perform sentiment classification from one domain to another. Studies proposed a method for transfer learning based on the multi-layer convolutional neural network (CNN). As a result, a CNN model was developed to extract features from a source domain dataset and to share weights in the convolutional and pooling layers between source and target domain samples. The model transfer from the source domain to the target domain and the fully connected layer's weights were fine-tuned. Therefore, retraining the network for the target domain is unnecessary. This method yielded relatively good results and demonstrates its ability to resolve the problem of the lack of labeled data in the target domain..

3.4.3 Multimodal sentiment analysis

Multimodal sentiment analysis is an expanding field of study. This field aims to include audio and visual data in the process of sentiment analysis, in addition to text. This is due to the fact that web 2.0 enabled people to express their emotions through images, videos, and audio files in addition to text. Thus, sentiment analysis and affective computing have evolved from unimodal analysis to more complex forms of multimodal analysis. A multimodal strategy may be bimodal, employing various combinations of two modalities, or trimodal, incorporating three modalities. However, the fusion of multimodal content or information presents a number of difficulties, including fusion strategy, hyper-parameter tuning, interpretability, and speed. To overcome this barrier a numerous information fusion techniques were developed. In addition, a number of studies utilizing information fusion for opinion mining have been examined.[18] Different feature fusion strategies were presented to improve the three-modal multimodal fusion mechanism. One such well-known strategy is explained below: That strategy hierarchically combined the feature vectors of various modalities, as opposed to simply concatenating them. It began by fusing modalities two in two, and then fused all three modalities. The findings indicated that this method reduced the error rate by 5 to 10 %. Several other works are being underdeveloped to incorporate this method into the field of sentiment analysis.

4. EVALUATION METRICS

Various metrics are utilized to evaluate the performance and efficacy of an approach or proposed model. This final step of developing a model is crucial because not all metrics are appropriate for a given problem, and sometimes a new evaluation metric can be introduced to evaluate the newly proposed approach. Measuring and comparing the performance and effectiveness of a model can be affected by the metrics selected. Confusion Matrix (also known as error matrix or truth table), Receiver Operator Characteristic (ROC), and Area Under the Curve (AUC) are among the techniques used to evaluate and summarise the performance of a classification model.[19] A basic confusion matrix is a 2 by 2 matrix, that summarises the number of correct and incorrect samples predicted by a classifier, where:• TP represents the number of positive samples that were correctly predicted as positive by the classifier;

• FP represents the number of negative samples for which the classifier incorrectly predicted positive results.
• FN is the number of positive samples that the classifier incorrectly predicted to be negative.
• TN represents the number of negative samples that the classifier correctly predicted to be negative.

A perfect classifier would have no FP or FN entries. Sadly, this is not the case in the real world, as every model has limitations that prevent it from being 100% accurate the majority of the time. In addition to or in lieu of the aforementioned metrics, additional metrics have been used to evaluate various approaches. ROC evaluation metric, AUC evaluation metric, Kappa evaluation metric, and Root Mean Square Error (RMSE) evaluation metric. However, obtaining a model with good performance

is not always straightforward, and it is sometimes necessary to solve certain issues during or before the training process, such as preventing overfitting or dealing with noisy data, particularly when using machine learning algorithms. L2 regularisation is a popular algorithm to prevent overfitting for the logistic regression classifier. Cross-Validation is another technique for preventing overfitting and underfitting (CV). Although accurately predicting sentiment is extremely important, it does not always provide complete information due to the subjective nature of emotions. Therefore stacked ensemble methods have been proposed for predicting the sentiment intensity level. Using multiple-layer perceptrons, they combine the results of three deep learning models LSTM, CNN, and GRU. Predicting the level of sentiment aids in comprehending the precise emotion associated with a given level of sentiment analysis.

5. SENTIMENT ANALYSIS CHALLENGES

5.1 Sarcasm detection

According to the Macmillan English Dictionary, sarcasm is the act of saying or writing the opposite of what someone means or speaking in a way that is intended to make another person feel stupid or demonstrate anger. As an example, sarcasm complicates the task of sentiment analysis when someone writes something positive when he actually means something negative, or vice versa. We use sarcastic expressions frequently in daily life. By automatically identifying sarcastic expressions within a given text, sarcasm detection is becoming increasingly popular as a solution to the problem of deceptive sentiments. Due to the ambiguity and complexity of sarcasm, sarcasm detection is a very difficult NLP task. [20] Deep learning was used to detect sarcasm in real-time in a mix of English and an Indian native language (Hinglish). The proposed model is a combination of bidirectional LSTM, a softmax attention layer, and a convolutional neural network. Softmax was used to learn the semantic context vector for English features from the GloVe word representation and forward it to CNN. The CNN model combined with the HindiSenti (Hindi SentiWordNet) feature vector and punctuation-based auxiliary features. With a higher classification accuracy of 92.71%, this model outperforms the baseline deep learning models.

5.2 Negation handling

Negation words, such as not, neither, nor, etc., can reversethe polarity of a given text, making their treatment crucial for sentiment analysis. For example, the sentence ”This movie is good.” is a positive sentence, while ”The movie is not good.” is a negative sentence. Unfortunately, in some approaches, negation words are eliminated because they are included on Stop-Word lists or are implicitly disregarded because they have a neutral sentiment value in a lexicon, which has no effect on the final polarity. However, it is not simple to accomplish this task by reversing the polarity because negation words can appear in a sentence without altering the text’s meaning. Lazib [21] proposed a syntactic path-based hybrid neural network for negation scope detection. The CNN model captures relevant syntactic features between the token and the cue along the shortest syntactic path in both constituency and dependency parse trees, whereas The Bi-LSTM learns the context representation along the sentence in both forward and backward directions. Their model earned an F-score of 90.82 percent.

5.3 Spam detection

In the field of sentiment analysis, spam detection plays a critical role. As online opinions influence consumer purchasing decisions, spam and fake reviews can harm the brand reputation and artificially manipulate users’ perceptions of products, services, companies, and other entities [22]. Developing a spam detection system capable of identifying fake reviews among a large number of reviews is a very difficult task due to the lack of observable differences between reviews. Among the systems proposed to perform the task of spam detection, Studies say that system efficiently employs three features: review sentiment and comments, content-based factor, and rating deviation. This method uses the comment data to determine whether a review is a spam or not. The authors classified the unlabeled data and two oversampling using the labeled data and a machine-learning model. Because the number of spam reviews is typically much lower than the number of genuine reviews, it is necessary to employ techniques for making the classes comparable. The F-score of their system was 91 percent.

5.4 Resolution of anaphora and coreference

Anaphora is a correlation between linguistic terms. Identifying what a pronoun refers to in a sentence is useful for sentiment analysis, particularly aspect-based analysis because it helps to extract all aspects of a given entity. Unfortunately, pronouns are typically disregarded or eliminated during preprocessing. Some research provided an exhaustive overview of the fields of coreference resolution and anaphora resolution, which are closely related. Studies proposed an Enhanced Anaphora Resolution Algorithm. This algorithm offers inter-sentential anaphora resolutions by identifying compound nouns and resolving the PoS for each word. The algorithm achieved superior performance in comparison to the conventional anaphora resolution technique.

5.5 Word sense disambiguation (WSD)

A word can have multiple meanings, and depending on the context and the domain used, the meaning of this word can vary depending on the situation. The objective of word sense disambiguation is to determine which sense of a word is employed in a given sentence. For instance, the word ”curved” refers to a positive context when describing a television, but may refer to a negative context when describing a mobile phone. Therefore, determining a word’s meaning from a sentence is extremely difficult. Wang [23] proposed a knowledge-based method that relies on the well-known lexicon WordNet to solve this challenging task. Using Latent Semantic Analysis (LSA) and PageRank, respectively, this method models the problem of WSD with semantic space and semantic path concealed behind a given sentence. The experimental results demonstrate the efficacy of this method, as it has produced favorable results. Word polarity disambiguation (WPD) is an additional challenging problem.

WPD aims to clarify the polarity of sentimentally ambiguous words in a given context. This issue was addressed using a Bayesian model and opinion-level features. They investigated the level context by identifying intra- and inter-opinion characteristics. The Bayesian model was used to improve the effectiveness of opinion-level features and to resolve polarity probabilistically.

5.6 Languages with few resources

The majority of research in the field of sentiment analysis has focused on the English language or other languages with an adequate amount of linguistic resources (e.g. sentiment lexicon and labeled text corpus). As stated previously, supervised learning techniques are the most commonly used for sentiment analysis. However, these approaches rely heavily on linguistic resources, which are expensive to acquire for less-common languages . Low-resource languages are the types of languages that suffer from a dearth of linguistic resources (or under-resourced languages). Several methods can be employed to circumvent this issue: constructing linguistic resources from scratch; employing unsupervised, semi-supervised, and transfer learning approaches, as described in Sections 4.1.2, 4.1.3, and 4.4.2, respectively. Studies proposed an approach to exploit the abundant English resources for the classification of Chinese sentiment. Two denoising autoencoder classifiers were trained in English and Chinese views, respectively. The final sentiment classification results were then determined by combining the two results in two views. This demonstrates the usefulness of cross-lingual sentiment classification approaches for low-resource languages. Some researchers proposed a graph-based semi-supervised method for document-level sentiment classification in languages with limited resources. The authors investigated the usefulness of two graph-based algorithms, label propagation (LP) and modified adsorption, with limited labeled data (MAD). These methods contribute to an increase in labeled instances, resulting in more training data for languages with limited resources.

5.7 Sentiment analysis of mixed-code data

The simultaneous usage of vocabulary and grammar from many languages is known as code-mixing (CM). It is fairly common in communities that speak many languages and presents a significant barrier to NLP activities like sentiment analysis. The discovery of compositional interpretation, which is necessary for performing sentiment analysis using rule- and machine-learning-related approaches, is hampered by the lack of grammar rules for code-mixed phrases. One of the biggest challenges is the lack of established mixing guidelines because mixing is dependent on the individual . New linguistic models are therefore necessary for sentiment classification on code-mixed data. The issue of linguistics modeling for the hybridized form of English and Hindi text . Their study shows that the main problem of the CM language model as well as the main reason why conventional models would perform poorly is switching points (locations when a person shifts to a different language). Among the few research to look into CM, is the research by Lal [24], even though it is a sizable obstacle. For the purpose of sentiment analysis on mixed English-Hindi coding data, the authors presented a hybrid architecture. They separated this design into three parts, each of which aimed to address a different issue. The suggested architecture can reach an efficiency of 83.54 per cent, according to tests on a database of code-mixed online posts.

VI. CONCLUSION

An outline of opinion mining as well as its associated approaches is given in this study. The main goal of this work is to evaluate and categorize the most popular classification methods used in sentiment analysis. Quick review methods of processing and the possible challenges were also discussed. Different methods for categorizing attitudes were categorized and their benefits and drawbacks were examined. The most often used method in this area is supervised machine supervised learning because of its simplicity and great accuracy. However, other techniques (such as reinforcement learning) offer a powerful solution to certain problems and challenges in the field, such as the lack of labeled data or other NLPrelated tasks. The difficulties presented later demonstrate that sentiment analysis remains an unexplored area of research. In this field, the English language is the most studied, but interest in other natural languages has increased in recent years. Still, resources for these languages are limited. Consequently, constructing useful resources for other natural languages besides English, such as building datasets and generating lexicons, can be an intriguing area of future research. The entire review revolved around the techniques used to sentiments expressed in the English language.

References

[1] KashfiaSail, RedaAlha”Emotion and sentiment analysis from Twitter text”Journal of Computational Science Volume 36, September 2019,101003

[2] M.V. M¨antyl¨a, D. Graziotin, M. Kuutila, The evolution of sentiment analysis—A review of research topics, venues, and top cited papers, Comput. Sci. Rev. 27 (2018) 16–32, https://doi.org/10.1016/j.cosrev.2017. 10.002.

[3] F. Hermitian, M.K. Sohrabi, A survey on classification techniques for opinion mining and sentiment analysis, Artif. Intell. Rev. 52 (2019) 1495–1545, https://doi.org/10.1007/s10462-017-9599-6.

[4] Marouane Birjali, Mohammed Kasri, Abderrahim Beni-Hssane”A comprehensive survey on sentiment analysis: Approaches, challenges, and trends”,2020

[5] Sibi Mathew, ”Review On Practical Applications of Social Media Data Mining”, International Journal of Scientific Engineering and Research (IJSER). , SE23108203033,2023

[6] 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.

[7] H. Sankar, V. Subramaniyaswamy, Investigating sentiment analysis using machine learning approach, in: 2017 Int. Conf. Intell. Sustain. Syst, IEEE, 2017, pp. 87–92, https://doi.org/10.1109/ISS1.2017.8389293.

[8] A. Jurek, M.D. Mulvenna, Y. Bi, Improved lexicon-based sentiment analysis for social media analytics, Secur. Inform. 4 (2015) 1–13,https://doi.org/ 10.1186/s13388-015-0024-x.

[9] 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.

[10] Ansari Fatima Anees,Arsalaan Shaikh,Arbaz Shaikh,Sufiyan Shaikh,"Survey Paper on Sentiment Analysis: Techniques and Challenges",January 15, 2020
[11] Jaspreet Singh, Gurvinder Singh & Rajinder Singh Optimization of sentiment analysis using machine learning classifiers December 2017
[12] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, eds., A Practical Guide To Sentiment Analysis, 5, 2017, pp. 1–196, https://doi.org/10.1007/978- 3-319-55394-8.
[13] P. Yang, Y. Chen, A survey on sentiment analysis by using machine learning methods, in: 2017 IEEE 2nd Inf. Technol. Networking, Electron. Autom. Control Conf, IEEE, 2017, pp. 117–121, https://doi.org/10.1109/ ITNEC.2017.8284920.
[14] Ashima Yadav & Dinesh Kumar Vishwakarma"Sentiment analysis using deep learning architectures: a review" pages4335–4385 (2020)
[15] Venkateswarlu Bonta1 , Nandhini Kumaresh2 and N. Janardhan3 ,A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 1-6 © The Research Publication, www.trp.org.in
[16] G. Yoo, J. Nam, A hybrid approach to sentiment analysis enhanced by sentiment lexicons and polarity-shifting devices, in K. Shirai (Ed.), 13th Work. Asian Lang. Resour, Miyazaki, Japan, 2018, pp. 21–28, https://hal. archives-ouvertes.fr/hal-01795217.
[17] Y. Al Amrani, M. Lazaar, K.E. El Kadiri, Random forest and support vector machine based hybrid approach to sentiment analysis, Procedia Comput. Sci. 127 (2018) 511–520, https://doi.org/10.1016/j.procs.2018.01.150.
[18] V. Vyas, V. Uma, Approaches To Sentiment Analysis on Product Reviews, 2019, pp. 15–30, https://doi.org/10.4018/978-1-5225-4999- 4.ch002.
[19] Mais Yasen; Sara Tedmori,"Movies Reviews Sentiment Analysis and Classification", 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)
[20] D. Jain, A. Kumar, G. Garg, Sarcasm detection in mashup language using soft-attention based bi-directional LSTM and feature-rich CNN, Appl. Soft Comput. 91 (2020) 1–11, https://doi.org/10.1016/j.asoc.2020.106198.
[21] L. Lazib, B. Qin, Y. Zhao, W. Zhang, T. Liu, A syntactic path-based hybrid neural network for negation scope detection, Front. Comput. Sci. 14 (2020) 84–94, https://doi.org/10.1007/s11704-018-7368-6.
[22] S. Saumya, J.P. Singh, Detection of spam reviews: a sentiment analysis approach, CSI Trans. ICT. 6 (2018) 137–148, https://doi.org/10.1007/ s40012-018-0193-0.
[23] Y. Wang, M. Wang, H. Fujita, Word sense disambiguation: A comprehensive knowledge exploitation framework, Knowledge-Based Syst. 190 (2020) 1–13, https://doi.org/10.1016/j.knosys.2019.105030.
[24] Y.K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing senti- ment from code-mixed text, in: Proc. 57th Annu. Meet. Assoc. Comput. Linguist. Student Res. Work, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 371–377, https://doi.org/10.18653/v1/ P19-2052.