SENTIMENT ANALYSIS ALGORITHMS AND ITS DIVERSE APPLICATIONS: A SURVEY


1,2,3,4Sona College of Technology, Department of Computer Science and Engineering, Salem, Tamil Nadu
1dvbharathi77@gmail.com, 2theetchenya@gmail.com, 3mari.btech@gmail.com, 4gviddyacse@gmail.com

ABSTRACT

In recent years, the field of sentiment analysis, which collects, analyses, and aggregates sentiment from text, has received a lot of interest. As this topic grew, it spawned a slew of other subfields, each targeting a different degree of inquiry. This article focuses on an overview of the most recent developments in this sector. In sentiment analysis, classic machine learning methods are employed to handle categorization problems successfully. Deep learning has been employed in various disciplines to get successful outcomes as a result of its rapid expansion. This review summarizes a number of deep learning models that have been employed in sentiment analysis applications. This assessment includes looks at recently developed techniques and various semantic analysis applications.

Keywords: Opinion mining, Sentiment Analysis, Classification, Evaluation metrics

I. INTRODUCTION

Opinions are vital in all aspects of human life since they are the primary mirrors of our actions. When individuals need to make a decision, they always want to know what other people think. Nowadays, businesses and organizations are always seeking client feedback on their products and services. Customers' reviews include a wealth of information that researchers may utilize to conduct study. The large number of viewpoints helps in decision-making. However, consumers may find it difficult to assess all web viewpoints owing to the vast quantity of viewpoints published on the internet. Because all reviews are provided in plain text in any natural language, we require assistance from other disciplines such as Natural Language Processing and Data Mining to extract useful information from them.

Opinion mining (also known as sentiment analysis) is the study of people's attitudes, feelings, and emotions concerning things like products, services, and their characteristics. Opinion mining, on the other hand, have different concepts, according to some experts. Opinion mining is the process of gathering and analyzing people's thoughts on a subject. Sentiment analysis, on the other hand, identifies and expresses sentiments as a text, which is subsequently examined. In general, sentiment analysis is explored at three levels: document level sentiment classification, sentence level sentiment classification, entity and aspect level sentiment classification, and entity and aspect level sentiment classification. The whole opinion about the document is categorized as positive or negative during document-level sentiment analysis. The whole document evaluation is treated as a single subject in this instance. Every sentence is classified as good, negative, or neutral in opinion mining at the sentence level. Sentences with no viewpoint or unrelated terms are referred to as neutral opinion. The document and sentence level evaluations do not accurately reflect the people's preferences. The goal of the feature or aspect level is to categorize sentiment in relation to entities and their attributes. Instead of focusing on linguistic structure such as clauses, sentences, or paragraphs, it focuses solely on opinion. Feature or Aspect Identification, Sentiment Classification, and Polarity Identification are the three main objectives of opinion mining. In general, we categorize all techniques into three groups: Aspect detection, sentiment analysis, and combination aspect detection and sentiment analysis approaches are all available.

Data was gathered via social media, product review websites, and other sources. This will help businesses make business decisions based on the outcomes of customer feedback on their products. Stock markets, news
stories, political discussions, and other areas may all benefit from semantic analysis. In general, there are two basic data collection strategies. The majority of product-related social media comments come from online forums, discussion boards, and consumer rating websites. Second, comments from Twitter, Plurk, and Facebook that aren't about the product. We have a different programme that retrieves the needed comments from the selected social media based on a certain characteristic. As a result, product analysis is simplified.

Many applications and improvements to semantic algorithms have been proposed in recent years. This study seeks to sort the current publications in this topic into categories according on the numerous SA methods, applications, and problems. All of the papers under consideration were published during the previous several years. For new researchers in this topic, this survey may be informative. It is the only survey that categorizes the numerous semantic analysis methodologies in a way that no other survey does. Opinions are classified according to their polarity, and dictionaries are used to construct a lexical corpus. A system's capacity to detect and apply information and abilities obtained in past activities to fresh activities or new domains that share some resemblance is known as transfer learning.

II. LITERATURE REVIEW

Many articles are summarized in Table 1. The summarization is divided into categories depending on the sentiment analysis technique, data sets utilized in the procedures, system outcomes, and scopes. It will be extremely valuable in assisting researchers in determining which topic of study to pursue in the future. The algorithm employed in the research article is specified in the second column. As a result, we may learn about the different methods that have been employed for feature selection, sentiment classification, and polarity identification in the last year. It will be beneficial to identify the many domains that researchers in SA may investigate for future study. The results of the algorithms utilized in the review papers are shown in the last column.

<table>
<thead>
<tr>
<th>Author &amp; Reference</th>
<th>Problem</th>
<th>Deep Learning Model Used</th>
<th>Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhai and Zhang [2]</td>
<td>For sentiment classification, to acquire an enhanced document vectors.</td>
<td>Semi-Supervised Autoencoder</td>
<td>By modifying the loss function, it learns a interpretation of the written texts.</td>
</tr>
<tr>
<td>Tang et al.[3]</td>
<td>To learn about representation of a document by taking into account the sentence connections.</td>
<td>CNN or LSTM from wordembeddings</td>
<td>Sentimental categorization is enhanced adaptively encoding the meanings of phrases and their intrinsic relationships in document representations.</td>
</tr>
<tr>
<td>Tang et al.[4]</td>
<td>In order to classify reviews, user and product representations must be identified.</td>
<td>LSTM</td>
<td>By capturing significant global indications like user preferences and overall product quality, superior text representations are generated.</td>
</tr>
<tr>
<td>Zhou et al.[5]</td>
<td>Document level classification of sentiments in Cross-lingual.</td>
<td>Attention-based LSTM network</td>
<td>The sentiment information is converted from high levelor global language (English) to a low resource or local language (Chinese) and aids in sentiment categorization.</td>
</tr>
<tr>
<td>Wang et al.[7]</td>
<td>To forecast the text's arousal ratings</td>
<td>CNN-LSTM model</td>
<td>Good accuracy is obtained in predicting the arousal ratings.</td>
</tr>
<tr>
<td>Guan et al.[8]</td>
<td>Polarity at sentence level</td>
<td>weakly-supervised CNN</td>
<td>Sentence representation that is only loosely driven by overall review scores is discovered, and sentence level labels are used to fine-tune it.</td>
</tr>
</tbody>
</table>

Table 1. Summarization of deep learning models developed in recent years
<table>
<thead>
<tr>
<th>Authors</th>
<th>Context</th>
<th>Model Use</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teng et al. [9]</td>
<td>To propose a Context-sensitive lexicon-based method</td>
<td>Bidirectional LSTM</td>
<td>The emotion strength, intensification, and negation of lexical feelings are identified successfully while building the polarity value of a phrase.</td>
</tr>
<tr>
<td>Dong et al. [10]</td>
<td>Twitter sentiment classification based on the target</td>
<td>AdaRNN</td>
<td>Depending on the context and grammatical organization of the tweets, the feelings of words towards the target is propagated.</td>
</tr>
<tr>
<td>Tang et al. [11]</td>
<td>To improve the LSTM model by considering the target into account.</td>
<td>Target-dependent LSTM (TD-LSTM) and Target-connection LSTM (TC-LSTM)</td>
<td>The target feature obtained is concatenated with the context data and utilized for aspect based polarity identification. The overall performance is enhanced.</td>
</tr>
<tr>
<td>Ma et al. [12]</td>
<td>Target and Context are considered for classification</td>
<td>Interactive Attention Network (IAN)</td>
<td>The significant words in the target expression are recognized by utilizing the combination of attention works. Also with the essential words in its whole context interactively</td>
</tr>
<tr>
<td>Chen et al. [13]</td>
<td>To identify the exact polarity in the complicated contexts.</td>
<td>Recurrent attention network</td>
<td>The recurrent attention network is trained and tested to identify the precise context of complicated reviews.</td>
</tr>
<tr>
<td>Tay et al. [14]</td>
<td>Dyadic interactions between aspect and context.</td>
<td>Dyadic Memory Network (DyMemNN)</td>
<td>The neural compositions for memory selection operations have been enhanced, which enhances the speed with which sentiment is identified.</td>
</tr>
<tr>
<td>Ponta, Let al. [16]</td>
<td>To enhance the approaches for locating financial-related reviews.</td>
<td>RNN with LSTM</td>
<td>The created models are capable of detecting emotions in market trends, neighbouring agents, and market propensity.</td>
</tr>
<tr>
<td>Wataru Souma et al. [17]</td>
<td>Enhanced news sentiment analysis</td>
<td>RNN with LSTM</td>
<td>The model classifies the news article as positive(negative) based on the observed stock price as positive(negative).</td>
</tr>
<tr>
<td>Hai Ha Do et al. [18]</td>
<td>ABSA</td>
<td>CNN,RNN</td>
<td>Product reviews and target-dependent tweets are classified effectively based on aspects.</td>
</tr>
<tr>
<td>P. Sasikala et al. [19]</td>
<td>SA of online products review</td>
<td>Deep learning Modified Neural Network (DL MNN)</td>
<td>Enhanced prediction of online review of the products is implemented significantly by Improved Adaptive Neuro-Fuzzy Inferences System (IANFIS), to thrash the problems in Traditional approaches.</td>
</tr>
<tr>
<td>Mohsen Ghorbani et al. [20]</td>
<td>On the Google cloud, to determine the polarity of words and do calculations on Google Colaboratory</td>
<td>ConvLSTMConv</td>
<td>An approach for assessing sentiments and categorising them into positive and negative categories.</td>
</tr>
</tbody>
</table>

The important classification of opinions are: Regular opinion and comparative opinion. Regular opinion is simply the normal opinion. Regular opinion is sub divided into Direct and indirect opinion. In direct opinion, the opinions are directly expressed on aspect or attribute. Eg Samsung has high Picture Quality’. In Indirect opinion, the opinions are not directly expressed on aspect or attribute. Many researches are going only in direct comments when compared with indirect one. The reason is, indirect opinions are difficult to identify. A comparative opinion conveys a relationship of similarities or differences between two or more entities, as well as the opinion holder's choice based on some of the entities' common characteristics. They have distinct syntactic forms as well as diverse semantic implications.

**Unsupervised learning approaches** uses the outcome of statistical analysis on a given corpus to recognize the opinion features. The various methods are

**Random Feature Selection**
As many characteristics as possible are chosen at random. Random selection isn't expected to perform well because it doesn't try to find useful features in the first place. When more functions are introduced, it might make the system noisier than it was before. As a result, each time a new randomly sorted list is created, it is impossible to reproduce our results.

**Collection Frequency (CF)**

CF is the total number of instances of a feature in the collection of reviews. It is merely a count and does not consider which documents or categories the characteristic appears in. A feature is counted twice if it appears twice in a document, and so forth. The characteristics are sorted by the number of times they appear in the document.

**Collection Frequency Inverse Document Frequency (CFIDF)**

Weighting the collection frequency values by the inverse document frequency for a feature yields the CFIDF. The idea behind this metric is that combining the local document frequency and the overall number of occurrences for a feature might result in a higher ranking than just the collection frequency.

**Document Frequency (DF) Threshold**

In the training set, the number of documents that possess a feature is tallied. This is done for each feature in the training set before deleting all features with a document frequency less than or equal to a specified threshold and features with a frequency greater than or equal to another threshold.

**Term Frequency Document Frequency (TFDF)**

A technique is proposed that uses the term frequency in conjunction with the document frequency threshold. They call it Term Frequency Document Frequency, and it has been proven to be superior to the DF threshold.

**Supervised machine learning methods** may be customised to function effectively in a certain domain, but unless transfer learning is used, the model must be retrained if it is applied to multiple domains. In addition, for model learning in any domain, a collection of labelled data is required. The various types of methods are

**Word Frequency**

It is the simple Word Frequency for a (feature, category) pair. As a result, it simply looks for proof of membership in a category. The global word frequency value for each feature is calculated by combining these values.

**Information Gain (IG)**

The main premise of IG is to determine how effectively each individual feature distinguishes the provided data set. The uncertainty of the feature (e.g. term) and the dataset is measured using information entropy.

**Mutual Information (MI)**

For binary challenges, Mutual Information may be demonstrated to be equivalent to Information Gain. However, the two are not comparable for multi-class issues (with global feature lists), as we demonstrate in this study (although rather similar). As a result, we provide Mutual Information as a distinct feature selection approach with its own equation.

**Chi Square (CHI)**

To evaluate the independence of two variables, feature selection by $\chi^2$ testing is utilised. The characteristics with the highest $\chi^2$ values for a category should perform the best when it comes to document categorization.

**Latent Semantic Indexing**

As a function of the initial collection of features, feature transformation algorithms produce a reduced collection of features. One of the most effective feature transformation approaches is LSI. The LSI approach converts text into a new axis system that is a direct mixture of the unique word properties.

### III. SENTIMENT CLASSIFICATION TECHNIQUES

The sentiment classification techniques are divided into machine learning approach and Lexicon based approach. They are
Supervised Learning
Probabilistic Classifiers
Linear Classifiers
Decision Tree Classifiers
Rule-based Classifiers
unsupervised learning
Meta classifiers

The lexicon based approaches uses lexicons which is a collection of sentiment terms. The categorizations are:

- Dictionary-based approach
- Corpus-based approach
- Statistical approach
- Semantic approach
- Lexicon based and NLP techniques

Probabilistic Classifiers

For classification, this classifier employs mixture models. Each class is assumed to be an element of the mixture in the mixture model. The chance of picking a specific term for that element is determined by a generative model of the mixture. There are three classifiers: Naive Bayes, Bayesian Network and Maximum Entropy (ME) classifiers. Naive Bayes (NB) is the simplest and most commonly used classifier. Bayesian Network (BN) assumes that all features are dependent. Bayesian belief work effectively, but it is costly. So that it is not frequently used. ME is a classifier that uses encoding to turn labelled feature sets into vectors.

Linear Classifiers

When the input feature is sparse, a linear classifier is one of the quickest classifiers and is employed in circumstances when classification speed is critical. Also, when the number of dimensions is big, linear classifiers perform well. Support vector machines (SVM) and neural networks are the most often used linear classifiers.

Decision Tree Classifiers

The nature of Decision Tree Classifiers is hierarchical. A decision tree is built. Now the decision tree may categorise the document by traversing the query structure from root to a specific leaf that reflects the document's classification goal.

Rule based classifiers

A collection of rules is used to model rule-based classifiers. The left hand side indicates a feature set condition, whereas the right hand side represents the class label.

Meta Classifiers

The meta classifiers is a classifier that does have its own classification algorithm, but the actual work is done by another classifier, it includes another processing step before the actual classification. As there is no algorithms that achieve the best accuracy in all situations, so we combine a set of learning classification algorithm into one system to improve the accuracy.

Lexicon- based Approaches

Opinion lexicon is a collection of opinion expressions and idioms. There are three primary methods for gathering opinion words. A collection of opinion words is manually gathered with their orientations in the Dictionary method. It helps to tackle the challenge of identifying opinion words with certain orientations and domains utilising a corpus in a corpus-based method. This is implemented using either statistical approach or semantic
approach. The statistical methodology seeks to find polarity values in a corpus by looking at the co-occurrence of adjectives. The semantic method is based on a number of concepts for determining word similarity and hence delivers direct sentiment ratings. NLP techniques are used with a lexicon-based method to uncover semantic correlations in the Lexicon–based and NLP method.

IV. EVALUATION METRICS

In sentiment analysis, the accuracy of the results is measured by the Precision, Recall, Accuracy and Error Rate and F1 measure. Consider the following scenario: we examined N documents using four cells designed to solve a binary classification issue. The cell contains the count for True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

Obviously, \( N = TP + FP + TN + FN \).

The Evaluation measures are:

- Precision = \( \frac{TP}{TP + FP} \)
- Recall = \( \frac{TP}{TP + FN} \)
- Accuracy = \( \frac{TP + TN}{N} \)
- Error = \( \frac{FP + FN}{N} \)
- \( F1 = 2*\text{Recall}*\text{Precision}/(\text{Recall} + \text{Precision}) \)

V. APPLICATION AREAS OF OPINION MINING AND SENTIMENT ANALYSIS

Opinion or feedback-based applications are becoming increasingly popular these days. Natural language processing is required since the Opinions are stated in their own language. The most important applications areas of Opinion mining and sentiment analysis are:

a. Purchasing Product or Service: Until recently, making a selection while purchasing a product or service has been a challenging undertaking for customers. As a result, they rely on the opinions of others via social media networks. People may quickly learn about other people's opinions and experiences regarding any product or service using this strategy. It is now simple to compare a product or service to different brands.

b. Quality Improvement in Product or service: The quality of a product or service is an essential requirement for a producer or manufacturer. Manufacturers may use sentiment analysis to obtain both positive and negative feedback on their product or service. As a result, they will be able to increase the quality of their product or service.

c. Marketing research: The result of sentiment analysis is either a good or negative view, which may be used in marketing studies. The newest trend of a product or service may be assessed using sentiment analysis tools. We may also deduce the cause for product market lagging from user evaluations, as well as why they have changed their minds over time.

d. Recommendation Systems: The method can express which opinions should be regarded and which should not be regarded by categorising people's opinions into positive and negative categories. We may use this information to complete product recommendations, new organisational schemes, and other concerns.

e. Spam detection: As long as the internet is available, anyone can post anything. This increased the likelihood of spam content on the internet. Spam material may lead individuals astray in their search for accurate opinions and feelings. Sentiment analysis can distinguish between spam and non-spam content on the internet.

f. Strategy Making: Management or the government can quickly assess new tactics by polling workers or the general population. Sentiment analysis allows us to use the analyzed data to create friendlier policies.
g. Decision Making: Sentiment analysis provides a summary of people's opinions that may be utilised to make decisions. The impact of the choice is assessed in order to determine whether there is anything new to learn or that will be useful in future decision-making.

h. Argument mapping software aids in the creation of logical connections between policy statements that are organised logically. The debate graph was created to offer a consistent framework to the large number of policy statements and to link arguments to supporting facts.

i. Voting Advise Applications help voters figure out which political party (or other voters) has done the greatest job throughout the reigning era (policy declarations) and which comes closest to their expectations.

j. The use of automated content analysis aids in the processing of large amounts of qualitative data. Management or the government can use a variety of instruments to swiftly analyse new strategies, such as polling employees or the broader public. Sentiment analysis helps us to build more friendly regulations based on the data we've analysed. This allows us to find related reviews and assign polarity to them.

VI. CHALLENGES AND ISSUES
The main challenges that are faced by Opinion mining and sentiment analysis are the following:

a. Language Problem: Almost everyone communicates in English across the world, and reviews are also written in the same language. This language has a large knowledge base, including dictionaries, lexicons, and corpora, which makes opinion mining straightforward. Researchers nowadays think in a new way and focus on languages other than English.

b. Fake Opinion: The false reviews that are created are not based on the reviewers' true experiences with the items or services, but are created for other reasons. They frequently contain insignificant good thoughts about certain items or services in order to promote the entities, as well as bad thoughts about certain entities in order to tarnish their reputations. One of the most difficult aspects of opinion mining is identifying bogus opinions.

c. Ambiguous Opinions due to sarcastic and ironic statements: In general, evaluations contain sarcastic and ironic language. The viewpoint of the text in a statement might be difficult to distinguish in sarcastic words, and we must apply multiple approaches to determine polarity. Irony is a sort of sarcasm that isn't quite as blunt as a sarcastic phrase. Both sarcasm and irony are certain to lead to erroneous orientation and opinion mining.

d. Detection of spam: Both genuine and spam information may be found on the internet. Spam content should be recognised and removed before the reviews are processed. As a result, sentiment categorization will be useful.

e. Least availability of opinion mining software: Opinion mining software is quite costly, and it is only utilised by big companies. Because the tools are not available to all users, not everyone can benefit from them. Melt water, Google Alerts, People Browser, Google Analytics, Hoot Suite, Tweet stats, Facebook Insights, and Page Lever are some of the tools that may be used to track sentiment research.

f. Integration of opinion with implied data: The feature and the opinion word are combined to make pairs to develop polarity as a consequence of the feelings. The inferred data dictate how sentiment or opinion words behave in real life.

g. Domain-independence: Opinion mining and sentiment analysis encounter significant challenges due to the domain-dependent nature of opinion terms. One set of characteristics may perform well in one area while performing poorly in another. We need to create a new sentiment analyzer for many domains.

h. Twitter –In our sentiment analysis, emoticons, abbreviations, absence of caps, bad spelling, punctuation, syntax, and word shortcuts provide a challenge.

VII. CONCLUSION
This survey article examines current advancements in the field of sentiment analysis and opinion mining. Table I is a comparison of current survey reports that shows the strengths and limitations of the algorithms employed in
sentiment analysis. It also goes through the numerous feature identification and sentiment classification algorithms that are often employed in this field, as well as which ones are the most efficient. As a result, Opinion Mining and Sentiment Analysis have a broad range of applications and are confronted with several research obstacles. Opinion Mining and Sentiment Analysis have become a hot topic in the natural language processing field, because to the rapid expansion of the internet and social media networks. More new and effective strategies are needed to solve the present obstacles that Opinion Mining and Sentiment Analysis are facing. The stated difficulties in this article have been addressed by a variety of techniques as stated thus far, but there is still a lot of room for development in this area.

REFERENCES