International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 7.328
Peer Reviewed Bi-Monthly
Vol.10 No.1 September – October 2022

AN ANALYSIS OF OPINION MINING ALGORITHMS

Krutikaben Chandrakant Patel[1] & Dr. Shabnam Sharma[2]

1Ph.D. Research Scholar, Department of Computer Science, Shri JJTU, Rajasthan, India

2Professor &Ph.D. Guide, Department of Computer Science, Shri JJTU, Rajasthan, India

Corresponding Author – Krutikaben Chandrakant Patel

DOI - 10.5281/zenodo.7890119

Abstract:

One of the most current challenges in the field of natural language processing is known as opinion mining or sentiment analysis (NLP). It is becoming more difficult for individuals to voice their opinions on various platforms such as Facebook, Twitter, and Yelp since the rate of technological advancement is accelerating at an exponential rate. The proliferation of social media has resulted in the creation of a significant quantity of new data, including comments, reviews, and opinions. On the other hand, the examination of this data is laborious and time consuming. As a result, the development of an intelligent system that can categorise or decide whether something is good, negative, or neutral is required. The purpose of this article is to provide a concise study and comparison of opinion mining with machine learning, deep learning, transfer learning, and the Hadoop framework. Specifically, the paper will focus on the topics.

***Keywords**: Opinion Mining, Machine Learning, Deep Learning, NLP Transfer learning and Hadoop*

Introduction:

A business analyst can better understand the social sentiment of their company's brand or product by using sentiment analysis, which is a form of contextual text mining. Sentiment analysis is used to mine text in order to extract subjective information from the material, and it provides assistance to business analysts in monitoring online data and conversations. Both the lexicon approach (LN) and the machine learning technique are viable options for accomplishing this goal. The machine learning strategy is simpler and more effective, but it requires labelled data in order to function properly [1]. The lexicon-based approach cannot compute the data if the term in question is not included in the dictionary. Sentence level, document level, aspect level, and user level analysis are the four types of sentimental analysis methodologies, according to [2].

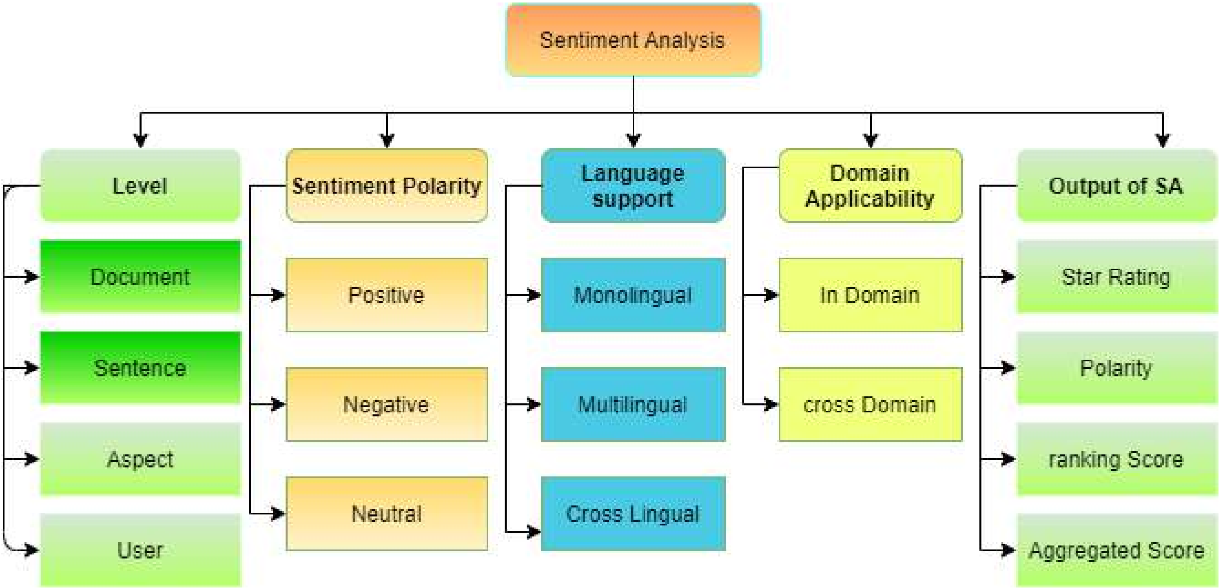

Fig.1: Overview of Sentiment Analysis

The thoughts expressed in each phrase are categorised as either good or negative using the sentence level approach. The primary objective is to ascertain the feelings conveyed by a single phrase. The document level approach groups the feelings expressed across an entire text into a single category. The strategy known as the aspect level concentrates on the characteristics of an entity. The user level is responsible for facilitating social interaction between the various parties.

In South Africa, the most frequent classification methods are based on machine learning and lexicons. Machine learning relies on testing and training to determine how to categorise data. The Lexicon approach is based on dictionaries that contain predefined positive and negative words [3]. The Lexicon approach is used as dictionaries that contain predefined positive and negative words [3]. While the second approach uses dictionaries that contain predefined positive and negative words [3]. Polarity is involved in analysing fine-grained sentiment, and the following categories may be used to categorise it: extremely positive, positive, neutral, negative, and very negative. After that, a rating score was derived from these categories; for instance, "extremely positive" mapped to 5 stars, whilst "very negative" mapped to 1 star. When dealing with several papers, one must first acquire each individual polarity before combining them into a single score.

Related Work:

In this part, we will provide a comprehensive explanation of the fundamental analytical technique and associated methodologies. The fundamental stages are shown in Figure 2 below.

ABasicFrameworkofSentimental AnalysisItisasequentialprocesshavingvarioussteps.

- **Input:**

The gathering of data is a vital component in the process of opinion mining. The data comes from a wide variety of sources, including Twitter, Facebook, online posts, blogs, microblogging sites, and review websites.

- **Preprocessing**

In preprocessing we need to process the collected data:

i) In the context of emotive analysis, brackets and numbers are completely meaningless. They should be eliminated since they are considered to be noise.

ii) Tokenization is a technique that may be used to a text in order to break it down into more manageable parts. such as "Removal of Extra spaces," "Emotions utilised replaced with their true meaning like Happy, Sad, abbreviation like OMG, WTF are replaced by their genuine meaning," "Pragmatics handling," such as "happyyyy as happy," "guddd as good," and "byeeeee as bye," etc. [4].

iii) Stop Word Removal refers to the process of removing words from an analysis that have no use, such as conjunctions (and, between), prepositions (a, an), and so on. [4] etc.

iv) The process of stemming involves removing postfixes from individual words, such as "ing," "tion," and so on.

v) utilized for the purpose of arranging the facts or the content.

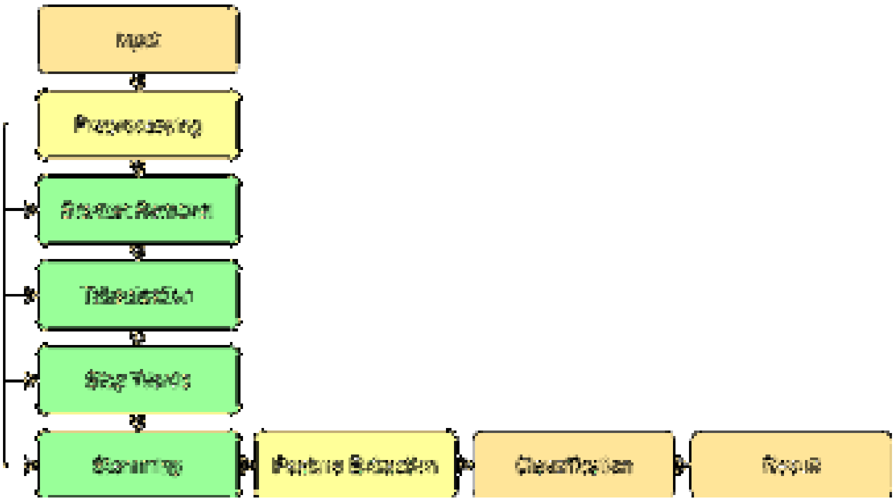

Fig.2:Basic framework of SA

- **Feature Extraction**

A feature is any observable aspect that may be assigned to a phenomena [1]. Using feature extraction techniques, we extracted various characteristics, such as adjectives, verbs, and nouns, and afterwards classified these features as having a positive, negative, or neutral polarity in relation to the data that was presented [5].

- **Classification**

In order to determine the polarity of user opinion, many different categorization techniques, many of which are well-known and widely used, are used.

Approaches for Sentimental Analysis:

Lexicon Based Approach: It is a strategy based on dictionaries, and it includes both positive and negative points of view. A positive score is given to the needed document if it contains a greater number of positive words; alternatively, a negative score is given to the document if it contains a significant number of negative terms. A neutral rating is assigned to the piece of writing in question if it has an equal number of positive and negative opinions. It is possible to construct and compile a dictionary-based method in a variety of different methods [5].

Dictionary based approach: Small amounts of words, known to have either a positive or negative orientation, are collected by hand and referred to as seeds [7]. doing a search for the synonyms and antonyms of the word in WordNet or another dictionary. Words discovered later are added to the seed list, and the iteration process starts over again. When there are no more new words to discover, the iteration process comes to a close.

Corpus Based Approach: The corpus-based strategy requires a large labelled dataset in addition to having dictionaries that are unique to the context in which they are being used [4].

Machine Learning Based approach: Computer programmes that can automatically improve their performance are the focus of the field of study known as machine learning. The primary goal of machine learning (ML) is to eliminate the need for any form of human involvement or aid in the learning process by giving computer systems the ability to learn on their own via experience. The computing approach is used by the algorithms so that they may learn directly from the data and not depend on a preconceived equation as a model [9]. Methods that are supervised and unsupervised are able to be distinguished in the context of sentiment analysis that is based on machine learning [6].

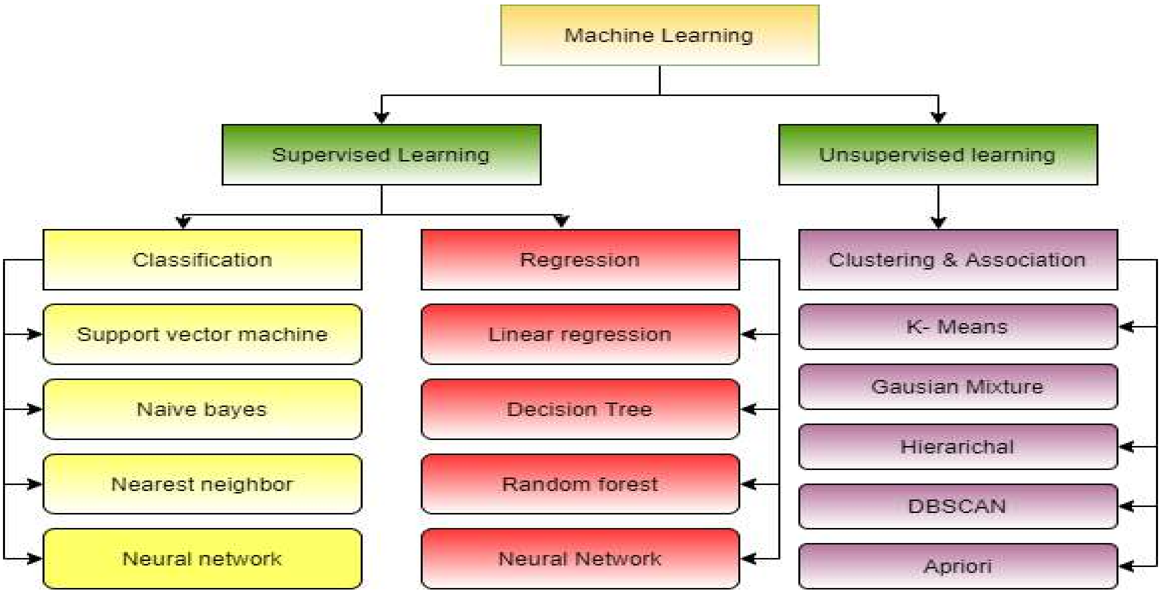

Fig. 3: Machine Learning Techniques for Sentimental Analysis

Deep Learning Based approach: Deep learning is a subcategory of machine learning that simulates how the human brain processes information by using computer programmes called artificial neural networks. It is a step up in sophistication. However, with deep learning, the neural network may learn to repair itself via its complex algorithm chain. When machine learning produces errors, human engagement is essential. Because of their inherent capacity for self-directed learning, deep learning models are very important. [17]

Through the use of deep learning models, sentiment analysis technologies are able to harness their full potential. It is able to be taught to interpret material beyond basic definitions, as well as the true sentiments and moods associated with the content. Several distinct types of deep learning models, including CNN (Convolutional Neural Networks), RecNN (Recursive Neural Networks), RNN (Recurrent Neural Networks), DBN (Deep Belief Networks), and HNN (Heterogeneous Neural Networks), may be used to carry out sentiment analysis (Hybrid Neural Networks).

Transfer Learning Based approach: Transfer learning (TL) [7] is a subfield of machine learning research that focuses on storing the information gained while solving one issue and applying it to another problem that is unrelated yet connected in some way. Because of this, there is no longer a need to train AI models from the ground up. The difficulty in training NLP models is increased by the absence of labelled data. Learning that can be transferred from one context to another is one of the most efficient approaches to addressing this challenge. Transfer learning has a lot of advantages, some of which are that it shortens the amount of time spent in training, increases the output accuracy, and requires less training data sets. When using deep learning, the initial few layers of the model are trained to recognise certain aspects of the issue. Therefore, in order to facilitate transfer learning, the last few layers of the trained network may be eliminated, and the network can then be retrained using newly added layers for the target task. Transfer learning is effective in a wide variety of computer vision applications, including pre-training on ImageNet, cancer subtype detection, digit recognition, building usage, and game playing. Text categorization is another application that has made use of transfer learning.

At the moment, the issue of negative transfer is one of the most serious obstacles that stand in the way of transfer learning. Transfer learning is only useful in situations where the beginning issue and the target problem are sufficiently comparable for the previous training to be applicable to the problem being learned.

Inductive Transfer Learning: Even if the source and destination domains are same, the tasks performed in each domain are not identical. The algorithms make an effort to make advantage of the inductive biases that are present in the source domain in order to assist better the target job. This may be further broken down into two groups, multitask learning and self-taught learning, depending on whether the source domain provides labelled data or not.

Transductive Transfer Learning: There are some parallels between the source tasks and the target tasks in this scenario. However, the source domain has a significant quantity of labelled data, while the target domain does not contain any labelled data.

Unsupervised Transfer Learning: It is analogous to inductive transfer, but the emphasis is placed on unsupervised activities inside the target area. The tasks are not the same, despite the fact that the source and destination domains are comparable. This labelled data cannot be found in either domain's repository.

Hadoop Based approach: Because more people are going online and participating in online commerce, the quantity of textual information that is available on the internet is increasing on a daily basis. This makes it more difficult to evaluate large amounts of data in an effective manner. Hadoop is a system that enables the gathering, storage, retrieval, administration, and distributed processing of massive amounts of data by using a cluster of computers and simple programming concepts. The Hadoop framework assists in the distribution of work over several clustering computers, which ultimately results in excellent performance. Additionally, each cluster has local storage and is able to carry out local processing. Hadoop is a development platform for parallel processing that is characterised by its high performance intensity, scalability, and flexibility.

Hadoop is capable of handling all elements of Big Data analysis for the purpose of determining sentiment. Hadoop's performance in the area of sentiment analysis may be increased by breaking the data up into modules, processing it on several computers, lowering the response time, and improving the system's failure tolerance by duplicating the data. It assists in the collecting of a wide range of unstructured data from numerous sources, in different forms, and across multiple domains, as well as the effective processing of this data in a multi-dimensional approach. When it comes to processing massive amounts of data, the accuracy of machine learning algorithms such as Naive Bayes improves significantly when they are implemented using MapReduce. The machine learning techniques that are given in Mahout work well with high-dimensional, huge volume, and complicated data, and they may be used in a variety of contexts. The Apache Open Source platform that makes use of Hadoop also offers applications with lower costs, which may be used to carry out sentiment analysis and assist organisations increase their profits.

Literature Survey:

In paper [9], an approach known as the semantic orientation calculator (SO-CAL) is introduced. This technique works on intensifier and negation. It was tested on datasets including movie reviews and obtained 76.37 percent accuracy overall. A lexicon-based technique is used in order to identify and categorise the feelings conveyed in a text. The research presented in employs a three-stage strategy to extracting sentiment, together with a document-level approach. In the initial step of the process, sentiment is first retrieved either automatically from datasets or directly from the internet. The positive and negative sets for this stage will be extracted from this dataset in the second step. In the third step, new document test sets are classified based on the lists that were acquired in the second stage. The F1 score for positivity is 0.717, and the

F1 score for negative record is 0.622.

The author of this research [11] use the method of sentiment classification in order to categorise evaluations of Chinese products. The foundation of their approach was an unsupervised categorization system that had the ability to teach itself by growing the vocabulary seed. It started out with a single word (good), which was described as having a constructive connotation. The original seed was iteratively retrained such that it could be used for sentiment categorization. After that, the ratio was determined by applying the criteria of opinion density to the data.

A Twitter opinion mining (TOM) framework is used in [12] for the purpose of tweet sentiment categorization. SentiWordNet analysis, emoticon analysis, and an improved polarity classifier are the three components that make up this hybrid methodology. The sparsity difficulties were alleviated by the suggested classifier, which used a variety of pre-processing techniques in addition to several sentiment analysis approaches. Tests using six datasets showed that the suggested method produced an average harmonic mean of 83.3 percent. The experiments were carried out to illustrate this.

The researcher in the study by [13] proposed a seven-layer structure to examine the feelings conveyed by phrases. CNN (Convolution Neural Network) and Word2vec are required for this framework in order to generate vector representation and semantic analysis, respectively. Word2vec is an idea that was presented by Google. In order to improve the validity of the suggested model as well as its generalizability, the Dropout technology, Normalization, and the Parametric Rectified Linear Unit (PReLU) have all been used. The framework was validated using the data set from rottentomatoes.com, which is a corpus of movie review excerpts. The dataset includes five labels: positive, slightly positive, neural, negative, and somewhat negative. The system was validated using this data set. The suggested model beat the prior models, such as the Matrix-Vector recursive neural network (MV-RNN) and the recursive neural network (RNN), with an accuracy of 45.5 percent. This was determined by comparing the new model with the previous models.

Conclusion:

The completion of traditional methods, such as those based on lexicons, might take a significant amount of time. They also have trouble generalising their findings to apply to other fields or industries. Even when using the most basic algorithms for machine learning, the procedures of feature engineering and feature extraction are the ones that consume the most time. Due to the fact that as the network gains knowledge, it is able to develop the features on its own, deep learning makes it possible to lessen the load of feature production. Training natural language processing models may be challenging when there is insufficient tagged data. Learning that can be transferred from one situation to another is one of the most efficient answers to this challenge. When it comes to training AI models, there is no need to begin from square one. Transfer learning provides a number of benefits, including a reduction in the amount of time needed for training, an increase in output accuracy, and a decreased resource need. It is possible to do sentiment analysis without compromising either the accuracy or the speed of the process. It is able to grow to bigger data sets while keeping the same level of speed. The machine learning methods that are implemented using Hadoop are easier to understand and more flexible, and they need less lines of code. [1] Implementing sentiment analysis using Hadoop results in a process that is less complicated, more readily expandable, and delivers great performance at a cheaper cost. Nevertheless, owing to the properties of big data represented by the 5 V model: volume, variety, velocity, variability, and veracity, such processes call for the use of more complex processing methods. training data. Hadoop is used because it is a piece of software that is both open-source and free to use, and it is efficient for storing and processing massive volumes of data in patterns that are dispersed over several nodes.

References:

1) Atiqur Rahman, Md. Sharif Hossen, "Sentiment Analysis on Movie Review Data Using Machine Learning Approach" International Conference on Bangla Speech and Language Processing (ICBSLP), 27-28September,2019.978-1-7281-5242-4/19 ©2019IEEE.
2) C.Tan, L.Lee, J.Tang, L.Jiang, M.Zhou, and P.Li,"User-level sentiment analysis incorporating social network, "In Proc. Of ICKDDM, IEEE, pp.1397-1405, 2011.
3) X.Fan, X.Li, F.Du, XinLi, MianWei, "Apply word vectors

for sentiment analysis of APP reviews," In Proc. of ICSI, IEEE, 2016.

4) Digital India, "In 2016 International Conference on Information Technology (InCITe) The Next Generation IT Summit.

5) Mitali Desai, A. Mehta," Techniques for Sentiment Analysis of Twitter Data: A Comprehensive Survey", International Conference on Computing, Communication and Automation (ICCCA2016), ISBN:978-1-5090-1666-2/16/$31.00©2016 IEEE.

6) Reshma Bhonde, Binita Bhagwat, Sayali Ingulkar, Apeksha Pande, "Sentiment Analysis Based on Dictionary Approach", International Journal of Emerging Engineering Research and Technology Volume3, Issue1, January 2015, PP51-55 ISSN 2349-4395 (Print) & ISSN2349-4409

7) Siva Kumar Pathuri, Dr. N. Anbazhagan, Dr. G. Balaji Prakash, "Feature Based Sentimental Analysis for Prediction of Mobile Reviews Using Hybrid Bag-Boost algorithm, IEEE 7th International Conference on Smart Structures and Systems ICSSS2020.

8) A.Harb, M.Plantié, G.Dray, M.Roche, F.Trousset, and P.Poncelet, "Web Opinion Mining: How to extract opinions from blogs?, "in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 2008: ACM, pp.211-217.

9) T.Zagibalov and J.Carroll, "Unsupervised classification of sentiment and objectivity in Chinese text, "in Proceedings of the Third International Joint Conference on Natural Language Processing: Volume- I,2008.