



# Social Media Mining Using Machine Learning Techniques as a Survey

Ashish A. Bhalariao<sup>1</sup>(✉), Bharat R. Naiknaware<sup>2</sup>, Ramesh R. Manza<sup>1</sup>,  
Vandana Bagal<sup>3</sup>, and Shobha K. Bawiskar<sup>4</sup>

<sup>1</sup> Department of CS and IT, Dr. B. A. M. University, Aurangabad, MH, India  
aashish.bharao@gmail.com

<sup>2</sup> Dr. G. Y. Pathrikar College of CS and IT, MGM University, Aurangabad, MH, India

<sup>3</sup> K. K. Wagh Institute of Engineering Education and Research, Nashik, MH, India

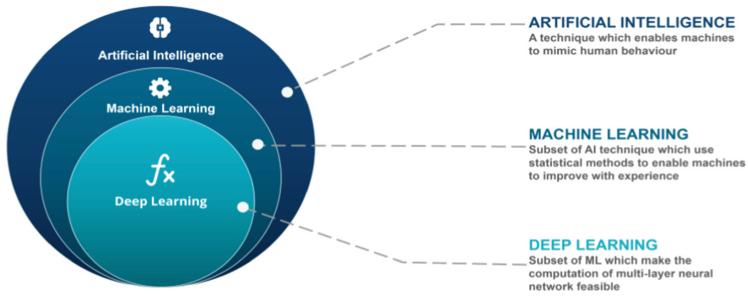
<sup>4</sup> Government Institute of Forensic Science, Aurangabad, MH, India

**Abstract.** In today's world an online existence and social media users utilize various social media platforms to express or comments their observations and opinions. The role of social media platforms are predicting Government Initiatives, Election results, product Analysis, business analysis, movie popularity, sports outcomes and stock market analysis. This review paper proposed the opinions are expressed through different social media platforms can be used for retrieving or extracting the real time predictions on several trends. As per the sentiment identification outcome find the features in the form of Positive (+ve), Negative (–ve) and Neutral (=). In this proposed research methodology, here collect user's reviews on particular trends, then preprocessed it, creation of the features and selecting for data classification using different machine learning classifiers and predict the result. For better performance, used advanced preprocessing techniques will be applied to cleaning the data. For Sentiment Classification will be used machine learning algorithms or techniques like (SVM) Support Vector machine, (ME) Maximum Entropy, (NB) Naïve Bayes and (DT) Decision tree. As per existing techniques, It is very difficult to mine the correct predictions from social media. Therefore, the prediction model will be designed for doing the prediction using real time data from Twitter. An opinion from text or comment posted on social media platforms by various categories of users is one of the critical and time consuming tasks in the field of opinion mining and analysis. The importance of this proposed intelligent system for social media is to automatically providing polarity from unstructured data in the form of text in English language for effective decision making.

**Keywords:** Social Media · Machine learning · Support Vector Machine · Clustering · Map Reduce · Deep learning

## 1 Introduction

Data is a very important aspect in the present and future also, it plays a very valuable and important role in decision making for end user, organization, Education sectors,



**Fig. 1.** (AI) Artificial Intelligence Vs (ML) Machine Learning Vs (DL) Deep Learning [29]

Government or Semi government sectors, Private sectors such as health, Pharmaceutical, product based industries, Commercial and many more. Social media grow to be much interesting and popular from the above mentioned sectors. This insights or information is regularly shared so services is definitely improves and overall costs is decrease using opinion which is generated by online user. All Data Mining and Knowledge or fact based mining become very popular because of user are process on the data and getting knowledge or insights of different area such as health, Social, etc. In Natural Language Processing, Sentiment Analysis or opinion mining is very significant application and now days it is globally approved and accepted by various outline of the organizations. The important task is to retrieve or extract insights from online existing data on social media and determine the impact of the social media as per different brands and various types of the aspects. All the commercial business sectors or some non profitable sectors such as Manufacturing, Pharmaceutical, Hospitals, Education institutions, Automobiles, E-Commercial sites etc., needs to extract useful Information from data which is available in digital form has been widely studied and very helpful to take decisions for business development and user satisfaction [1, 15] (Fig. 1).

To carry out the sentiment on reviews, opinions, user data and for classification of the text there are two main categories like classifier based approach where Machine learning approach have been studied and next is lexicon which is dictionary based method indicates positive, negative and neutral sentiment or opinion lexicons of words helps to find out sentiment of users text or data. ML is a division of AI and computer science which indicates on the utilized of data as well algorithms to represent the way so as to humans or users learn, slowly enhancing its accuracy.

### 1.1 Machine Learning (ML)

Machine learning is a very effective aspect of the advanced area of data science. All the way through the reason of statistical methods and algorithm techniques are trained to make the judgment or classification, discovery important insights inside data mining development. These important insights afterward compel decision making inside the business as well application [2, 18].

There are three primary categories of Machine learning classifiers as follows;

### **Supervised Machine Learning**

It is described by the making use of label (tag) datasets to teach the algorithms techniques to organize data and forecast result precisely. As an input text is submitted intensely on the model, amends it is weights awaiting the model have been matched properly. This happens as piece of the irritable justification. Supervised learning aims to resolve for general real-world problems, like classifying spam in different folder from main inbox. A few methods used in supervised include NB, NT, logistic regression, linear regression, SVM, random forest and more [3]. Supervised learning can be divided into two forms of tribulations: classification as well regression:

Classification tribulations indicate a technique to dispense data into exact type, like differentiating chairs and tables. It may be used to categorize spam in a diverse folder from inbox. SVM, Linear classifiers, random forest and decision trees are all general forms of classification techniques. Regression be one more type of supervised model that utilizes algorithm to recognize the association among different variables. This model is beneficial for analyzing numbered inputs depend on various data end, like sales profits prediction for organization.

### **Unsupervised Machine Learning**

Unsupervised learning, also well-known as unsupervised machine learning, employs machine learning algorithms to analyze or examine and cluster unlabeled datasets. These types of algorithms find out data grouping or hidden patterns without the interest for human intervention. Its capability to find out differences and similarities in information build it the perfect solution for exploratory data analysis, customer segmentation, cross-selling strategies, images and pattern recognition. It also utilize to minimize the number of features within a model from end to end the process of dimensionality reduction; singular value decomposition (SVD) and principal component analysis (PCA) are two general approaches for this. Another algorithms utilize in unsupervised learning comprise k-means clustering, neural networks, probabilistic clustering methods.

These models are used for 3 important tasks such as, union or Association, Clustering and Dimension reduction: Association is one more important type of unsupervised method that utilizes various system to search associations among values within related dataset dictionaries. These type of methods are regularly employ for examination and suggestion pop up messages, alongside the text of “Users Who seen this Also try for this” pop up messages. Second task is a data mining method for alliance untagged data depends on differences and similarities. This method is beneficial for image compression, etc.

Dimensional reduction is a method used whiles the integer of some features and dimensions in a mentioned big size of data. It minimizes the number of various data key to a controllable extent as well protecting the data reliability. Frequently, this is a method helps in the pre-processing of the data phase,

### **Reinforcement Learning**

Reinforcement learning presented a convenient means among both form of learning.

Throughout, it employs slighter tagged data to mentoring feature extraction, classification as of a bigger, untagged data. Reinforcement technique be able to resolve this issue of sufficient labeled data (or not being able to have enough money to tagged data) to teach a supervised techniques [2, 3, 14, 16].

## 1.2 Comparative Study Between Supervised and Unsupervised Methods

The major difference involving these two methods is the use of labeled/tagged datasets. In supervised method utilize labeled/tagged I/O data, as an unsupervised learning technique not utilizes labeled/tagged data. In supervised method, suggested algorithm be trained from the training dataset by creation a judgment as per the texts and regulates output. In supervised learning models be likely added correct than the unsupervised techniques, it involve straight user interference to tagged the data duly. For example, in supervised learning technique we can guess how long your journey will be based on the time of day, climate situation and many more. Other than at initial, we can train it to identify driving time may be extended due to climate situations.

Unsupervised learning techniques, it is work on their self to identify the scenario of the data. Here some user interference is required to validate or identifying some output figures. However, a data analyst would require validating that it builds sense for a recommendation engine to group kids clothes with an order of diapers, sauce and sippy cups.

Deep Learning is a sub branch of ML concerned to algorithm techniques stimulated as per function as well structure knows as neural network, system model study to carry out categorization job from data, sound, images. This model can attain very correctness, sometimes beyond human-level performance. Models are studied by putting a huge group or set of tagged data and neural contain many layers. Large amounts of labeled data required in Deep learning. For example, intelligent robot development required thousands of hours of video and millions of different images [3, 4, 14, 17].

## 1.3 Comparative Study Between Machine Learning (ML) and Deep Learning (DL)

DL is a particular form of machine learning. Working flow of ML start with related aspects manually retrieved from data such as images, aspects are then utilized to create a replica that alignment the objects in the image. Relevant features are routinely retrieved from images in the DL workflow. In addition, DL executes end to end training – where a model is decided rough data and a target to be execute, such as categorization.

### Deep Learning Models – Steps for Creation and Training

The three the majority ordinary ways community use DL to carry out classification are:

**Graze Preparation:** To prepare a network from graze, you collect a very huge tagged data set and intend architecture so as to resolve study the aspects and model. It is superior for latest projects, or huge number of output categories applications. This is a not as much of regular techniques, it means the huge amount of datasets and time of learning, these system classically requires couples of days to learning.

**Transmit Approach:** Mainly DL projects used a procedure that includes a pre trained model. It is begin with an accessible network, like GoogleNet or AlexNet, as well provide for in latest textual data including earlier unknown classes. Next to perform a new task, like categorizing only Shirts or pants instead of one thousand various other aspects.

**Aspect Retrieving:** The mainly important here to DL is use the featrure extractor network suitable for all the layers with understanding the tasked and features from photos. In training process all the features will be dragged out and we can use as an input to ML model like SVM [2–4, 14, 16, 17].

## 1.4 Types and Sources of Social Media

In general, there are various types of Social Media as per the user usage discussed as follows;

**Micro-blogging:** A blog is a dialogue or informal-adapted communication frequently represented within text, jpg, mpg, mp3 visual types on the (WWW) World Wide Web. Though, micro blog is an accurate type of a blog it permits minor data or text to be distribute online as well evaluate to a general blog. It allows a end users post slightly pleased possibly, small sentences, photos, small videos, or related to web pages. In the way to keep the blog in well-organized, bloggers generally utilize an amount of services like mailing or messaging. These type of services known as micro-posts where carried these to set the blog is known as micro blogging. Social media websites such as Twitter, Weibo, Facebook as well provides micro blogging services like tweets or status updating a few micro blogging services permit end users to be control user privacy settings who can comment and analyze on their micro blogs. Micro-blogging applications extend to huge form of areas such as advertising, e-commerce, product searching and e-marketing.

**Friendliness:** Two or more people associated through both by using online standard to make a socially known as friendliness network. Peoples are investing a lot of time being interact with each others on various social media platforms. The friendliness network users focus on particular or different topics also maintain the linking between them These network considered as friendliness network. For example, in an educational association, the students and researchers are actors or may be colleagues or friends, which having associations as friendships, and workshop or conferences are events. Furthermore, the regularly remarkable online platforms include Facebook, orkut or any other social media sites.

**Professional Corporate:** A professional or corporate related network or commercial service which is a group of social-network services to determine only on transportation and officially connected or professional character. Some important professional platforms include Xing, LinkedIn and Data.com Connect [5, 15].

**Images:** Image sharing platforms permit users to share images online locally or globally. Sharing indicate to only viewing the photos but not mandatory to download them. Impartially, these websites permits the users to backup photos, make the photo searchable, share photos, and even provide them the control, who can observe their shared photos. These website platforms could be also used for multipurpose such as image resourcing and image repository, visual literacy, academic and research purposes. Some considerable image sharing platforms include Flickr, Instagram, Pinterest and Flickr.

**Videos:** Online users are interested to share videos locally as well globally. Some of the video sharing web platforms permits users to attach small videos, though, others permits lengthy video content. Impartially, these websites permits users to store, share or comment, and manage viewership of shared data content. The web platform for Social Media are one of the important and useful to share video data. Live streaming is presently important features on social platform to solely acquire associated by other web users. Furthermore, it allows online users to get connect their existing Social Media logins like Facebook with website having the facility of video- sharing to immediately share videos. Some extensively these type of content sharing sites includes Vimeo, YouTube and Veoh [6, 21].

**Question Answer (Q/A) Forums:** On these types of forums, Online Users are easily to check the questions raised by different users with post the answers typically by specifically area expertise. Question/Answer discussion are frequently compound through enormous, expert business as well tending is to be useful as community for online users in the same areas converse questions offered by domain expert suggestions. In social Question/Answer services, anyone can ask a question as well comment or post an answer. These type of services followed into three different categories such as Digital reference where online user can request for specific assist from librarians. Second are the professional services where professional organization provide Question/Answer services and expert are responsible to give the answer of the question. The third one is the social or community Question/Answer, where everyone can raise and provide the answer. Distribution of the information in excess of online/live communities supports to users to promote self-presentation, examine appreciation and social online learning a few considerable Question/Answer forums include Quora, Stack Exchange [7, 23].

**Social Media Analysis** is the web-based technique applies to make over the online communications approved in excess of virtual networks with communities. The 2.0 interactive World Wide Web based projects are collected of content generated with the user, like text post, comments, data and videos produced from end to end online transportation. Generally, user generated service-especially Social Media are managed through a number of associations. Social Media Analysis is the approaches to gathered data from various Social Media platforms, like Facebook, Twitter etc., then access them with evaluate facts in the direction to take important decisions. Considerably, the fact be modifying, updating and expanding, to appreciate live user practice, intents, insights and sentiments [8, 10, 15, 24].

## 2 Survey of Literature

Social media turned into greatly accepted as of biomedical and health care information. This data is easily available for users thus health care services be get better as well patient treatment cost can be reduced with the use of sentiments received from the end users. He has suggested structure on drugs side effects as well negative, positive comments from users toward obtain improved health care. In this regard, some textual data is available so that these medical related documents are regularly helpful.

These medical records contain useful data about symptoms as well related medicines or prescription. Users can acquire various ideas by referring these sorted data so it is

very useful to fetching data using available datasets. In this Paper, user opinions are measured to search better medications and find out the medicine side-effects. He has concentrate on the cancer patient symptoms. Here, expert doctors list of proposals used for cancer medicines as per the symptoms and get it available all the medicine list and better treatments for the users on this forum. He has enlarged his study in social media addicted to Knowledge mining and data mining as well collects users' views resting on different medicines for cancer. Associated datasets are gathered using forums of cancer site. All these data is in tweets of users which is associated with treatments of cancer and its types. To solve the problem of big size of datasets, it is divided into the small parts. Every part of these paragraphs is to be allowing for as separate or divide document a few phrases be extracting from positive text of document and this phrases be extracting within every document. Later than discover positive, negative decision of every user, determine the (SFR) satisfaction and (DFR) dissatisfaction ratio of the user thus at the end of this, final result is display about users feedback in terms of positive and negative. For this, construct a system to retrieve side effect of treatments, patient symptom or medicines names by referring available data at forum in the form of semi-structural, and unstructured Data. The overall classification of treatment, positive and negative impact, name of Medicines, name of Symptoms, various techniques of data mining, maximum pattern mining sequential and frequent text mining. Though, by using this research in the field of text or data mining is not easy and unproductive. Finally he has provided the result in the form of Patient satisfactory and dissatisfactory report. The highest SFR is 70% and DSFR is 60% [1].

In sight of the accessible difficulty in conventional recommendation technique of minimum effectiveness and correctness, this paper presented a machine learning based recommendation algorithm for social media. The technique of this algorithm is foundation on conventional personalized filtering algorithm as well combined with related features of social media users also the algorithm considers the rating aspects, enhancing efficiency as referring clustering algorithm. Finally, algorithm will be appreciating popular platform as Hadoop cloud. This research is relying on filtering recommendation methods of traditional collaborative. It gives contribution in the four aspects as follows;

- I. Retrieve user score data using user text of micro blog, join association among consumer and relationship among users. Finally, in this research revise the score rating by referring same online users information as well user relations, after that bung a element of the absent ratings..
- II. Excavate added aspect score rating by referring network as well increase the thing thus additional aspect will be suggested for the user. This research, recommend a technique to group users with rating content.
- III. It uses Map-Reduce programming model with combination of Hadoop and recommendations system to manage big size of dataset available on social media network. It will help to increase the speed of execution of algorithm.

In this paper, He has acquired source of data as micro-blogs Sina available on social networking platform. To gather data, Sina platform for API of Java version benefited. User blog data, user ID, user relational data are required as per the requirement to development process of recommend system. For that reason, as the initial user set randomly



micro-blogs data in the number of 50 are elected. As per the association between users, new users will be added into the previous user set. Douban net platform is used to crawl the data, there are 2000 films included initially and preferred as per the user popularity. To improve the efficiency and accuracy level, a social media recommendation algorithm based on machine learning have be firstly projected. User rating matrix as well constructed. In this research work, some limitations are found such as star effects, multiple connection not considered as well suitable clustering method will be useful because it is a time consuming process [13, 21].

The present period of 4.0 Industry, the WWW digital earth have a property of information or data, such as IOT, cyber security, mobile generated, industry, corporate, health sector, social media data, etc. Automated apps and intelligent systems are developed by analyzing these data as well the knowledge of AI and ML is required. There are different kinds of ML i.e. machine learning algorithms exist as well deep learning methods, are smartly use large amount of. In this research, a complete outlook based on these algorithms and used to improve the abilities and intelligence of a system. Here, real world application areas are explaining with the use of machine learning algorithms. Various domains are analyzed like smart cities, agriculture, cyber security, etc. as well highlighted research information and some other challenges in this area finally this research work provides good information for academia and industry professionals to make a real world applications. A booming ML based model and data the learning algorithms to be trained using collected live data and information associated with application as well help through smart result-creation. Based on methods of ML techniques, different application demand sectors and its importance are decorated. Thus based on discussed challenges, some research opening in particular domain which have to be focused by referring correct outcome in different domains of various application [3, 19].

Deep learning (DL) is an important for processing command in various aspects such as image, texts, speech processing. To be globally accession of internet and well in organize to development, of social media (SM), it is very difficult to analyzing live data with conventional techniques. For these type of problems Deep Learning is attractive solution. Here, author has discussed the architecture of Deep Learning using taxonomy wise outline, for social media analysis author highlighted various social media analysis based problems and deep learning oriented explanation. A detailed depiction of varied SM domain provided in this paper. Techniques of deep learning have very important features for data representation using different domain such as sentiment analysis, business, anomaly detection, user behavior as well a lot of.

Though, resource requirement to manage the big size of data and enhance the efficiency as well minimize the cost of computational processing, learning of data from mixed social data sources, DL-based methods provides the solutions. Use of deep learning challenges to be focused in a canonical way that helps to proves to be an edge for the scientific community [4, 17].

In this paper, it shows the comparison of intelligent algorithms, which are develop for OM in SM data from the last five years. The results indicate that using Support Vector Machine with part of speech or POS, Unigram and Bigram with J-48 achieve sentiment classification accuracy 92%. From this paper, few challenges are shown that require future next consideration for this significant topic such as Data Sparsely, there



are lot of extremely few useful labels in the training dataset that is utilized for Sentiment Classification. Preprocessing the rough data consumes long time to translate the required data around the world into a structured format. For example, Twitter and Facebook content have mistakes in spelling, abbreviations colloquial expression, lack of grammar, lack of punctuation and the unfinished sentences. All these are create the critical research challenge to manage this type of content [5].

People have on track stock investments is a good choice now a days. Through practiced direction as well smart practice, from stock return, people can increase their annual income. Sometimes Investment or stock is a quite risky and depend on the market situation. Excessive gross of asset experts joined through universal lack of knowledge be relevant to the economic things linking to the people, avoid several as of dealing in stocks. The fear of crash too works as prevention toward a lot of. This information boosted for manage the importance of machine learning to guess stock movement. By using the Twitter API, tweets are gathered and adopted sentiment analysis techniques. Various stock and its closing values data is to be processed for make an application which is used to predict price change, companies and its stock details. Such a forecast would to a great extent assist possible investor stock with winning knowledgeable result which would straightly involve to the profits. Twitter provides robust API for collection of the tweets. Two ways are there for this, search and streaming API. in the direction of conquer the limitation of API streaming, API search is used. The search API is REST API which permits to ask for exact fresh tweet queries. In forecasting changing a stock values, he has performed, SVM technique to show greatly competent, realistic model by Machine Learning techniques. Classification of positive, negative and neutral tweets give a high-quality outline of public humor. To combine large amount of data as well load it in the live form, Cloud services are good option here and as per the requirement retrieve the data from REST API. Additional issue related with the public mood also be analyzed. Here, author analyzed and studied several different methodologies such as, GPOMS provided exact people mood with six aspect as relation. Regularly stock changing price is predictable which is having 87.6% accuracy. Some surveys say that Opinion mining is reasonable. SVM technique shows 64.10% accuracy, The speedy method and better accuracy by using K-ELM. To be consideration news article summary, experimental outcome carried out. It shows separately the result of summary and headlined of news. Experiment result conducted and results were derived, independently for headlines and the summary of news articles [6, 22].

Author has applied intelligent neural network and fuzzy to analyzing airline company services. In this work, U.S airlines dataset are used. To create a model passenger tweets are used with study of various features tweet time, location, id, user comments etc. Python is used to create a model. Total 14610 tweets are collected from Kaggle. Various preprocessing techniques are also used on these retrieved tweets to determine the classification as positive, negative and neutral. Those tweets are not informative means not provide the positive or negative result, it is eliminated from dictionary. To find frequency of word occurrences, bag of words is used and tags in tweets are eliminated. The final results show using this technique with dataset provides superior accuracy rate which is more than 70%. The neural network based on fuzzy classifies positive and negative tweets to determine the opinion about flight services and hospitality in terms of good

or bad. Also estimate and compared accuracy patterns with SVM and RBF. Finally, it shows that accuracy level of IDNN is best compared to SVM and RBF. SVM and RBF is more suitable for small size of dataset. Here in IDNN fuzzy measures used and it is suitable to handle large dataset [7].

Online users post hundred million messages in a day on Twitter, and tentatively ten million from them are using geo-coordinates. This is very high demanding on visual interface and scalable large amount of data. This paper presented an outcome based on social media visual analytics for situation awareness. The 2013 German Flood heavily affected many regions in the southern and eastern parts of Germany. Author has collected Twitter text messages written in Germany from June 1 and 10. All the retrieved tweets are about German Flood in 2013 Crisis intelligence finalized the tweets by analyzing tweets with scatter-blogs, as well data is verified with true events with the help of tools and techniques. The final outcome of a system provides additional influenced of its practicability, manageability to handle these crisis. He has abstracted so as to managing by these type of aspects ready to users familiar about various challenges to handle these type of data as well facilitate them a knowledge about TagMap and advanced classifier which is better for visual analytics of social media. The main task is to be concentrate on the system have to be manage in environment as well ready on conventions in specific area. Advanced technique which is good for analysis shortens the evolution procedure as well supports for easily clear the concepts. Additionally, the main issue is privacy which is not a engineering problem and each and every aspect related with privacy is to be considered in all designing aspects. Techniques or models to be constructed on gathered information are intended as per the requirement by analyzing some important social media aspect such as users timelines, relation to different people, its outline data, affinity toward positive subjects, patterns association, and past details present within different method might dissimilar post of wrong users. In general attitude was that social media analytics tools would help them to manage with the data volumes, identify true situation reports, and enter the communication with people on a broader scale [8, 25].

To Mining the information related to traffic using social media is a very high demanding as per the current need of social media. Author has focused to retrieve traffic related microblogs from Chinese micro blogging platform that is SinaWeibo, Firstly, to study word embedding illustration which is a three billion micro blogs dataset, he has applied continuous bag-of word model. Measure up to the usual one-hot vector illustration of words, word embedding can imprison semantic similarity among words as well it is effective in NLP domain. The combination of convolution neural network model (CNN) and long short term memory (LSTM) used to fetched traffic related micro-blogs by providing the input parameter of learned word embeddings. To retrieved traffic data includes data acquisition by crawling Sina Weibo, word segmentation, classification of microblogs, word embedding. He has compared SVM technique with the projected techniques, the multi-layer perception model stand on word vector features and the SVM based on word vector features, and adopted the deep learning techniques. In this paper, data is crawling by using Sina Weibo website. This is free but more complete compared to requesting official APIs. The dataset for classifying traffic microblogs into traffic irrelevant and relevant was gathered by finding Sina Weibo with key words that are manually selected.

Finally he has collected nearly 40,000 candidate microblogs by finding Sina Weibo with key words. Here, three types of deep learning models are tested as CNN, LSTM used for traffic prediction and the combination of CNN-LSTM for microblog text classification. The detection results of traffic relevant and irrelevant categories with various classification approaches are determined. CNN-LSTM model reach the highest F1 score. The F1 measures of the LSTM, CNN and CNN-LSTM model are close for both traffic irrelevant and relevant class. All these different three models proposed in this paper enhance the performance than that of the bow-SVM model which shows the good performance, the SVM model and the MLP model. The LSTM model has the maximum recall value for traffic relevant class, yet a moderately low precision. A poorer precision value means that a model has less accuracy and has more false positive instances. For traffic relevant class, the precision of the bow-SVM model is the maximum compared to that of other methods, other than this model has a minor recall which means it would notice fewer traffic relevant microblogs and therefore may miss out few traffic detail information like traffic incidents [9, 26, 27].

Author has focused to know the basics of opinion analysis; methods utilized for opinion analysis as well opinion extraction. Drawbacks, accuracy, advantages of these techniques are also compared. Various techniques and algorithms are used to obtain opinion analysis. Silently many researchers are determined for development of previous methods and developing latest effective techniques. Using Machine learning the performance of current task using previous experience is improved. Machine learning methods first train the algorithm with few particular inputs parameters to formulate a model. Presently on this model is utilized to test the latest data for classification. Some of the methods are SVM, Decision tree, Naïve Bayes, Rule based approach. These three important techniques utilized in sentimental analysis are studied and analyzed based on their accuracy and performance. As huge sentiment data sources are available, it is being utilized in various applications. Some important applications of SA such as Politics, Business, Summarization, Recommender system, Government intelligence etc. among this Sentiment analysis or opinion mining is facing a lot of challenges such as Domain generalization, Negation Handling, Pronoun Resolution, Language generalization, Detection of spam, World knowledge, Single sentence possessing multiple opinions, Synonym gathering, Comparative sentence, etc. [10].

Author has presented a scalable model for real-time sentiment analysis using Twitter data. The proposed system based on feature extraction from tweet texts, Using semantic information and morphological features, feature are extracted from tweet text. To train the different classifiers supervised learning method is adopted. For the retrieval of important features, a labeled tweets dataset is preprocessed. Tweets are collected from Twitter stream; features are retrieved after the preprocessing task. These retrieved tweets are considered as an input to classifier. In this work, various statics are calculated, classification and preprocessing done using Python tool. The final implemented model provides the facility of classification, feature extraction and capable to handle huge input data. With all the considered datasets similar trends are observed. The work shows that Part-Of- Speech tags, prior polarity as well Emoticons are the beneficial features in the analysis of Twitter data. Here two annotated manually tweet dataset, 1.6M tweets8, 2013 SemEval Task 2, 2015 SemEval Task 11 are experimented. By implementing the Storm's

Multilanguage protocol platform python scripts are run using Java. He has performed experiments on a Okeanos7 cloud platform which is an IAAS service for the Academic Community and Greek Research. This is done with Ubuntu LTS and installed Apache Storm with six virtual machine. With consideration of accuracy, MSE, cosine the result of MaxVote and Linear SVM are well performed [11].

In this paper, developed a sentiment analysis from product reviews using hybrid approach. This model classifies or categorizes the product review as positive or negative as per the query posted in the concern product reviews accessible by the amazon. Data taken from the "Amazon" which have 1000 instances divided into positive (500) and negative (500). In this paper, Cross Validation method with fold value equal to 10 has been used for training and testing phases. Using Random forest technique, we see that 810 reviews are correctly classified among 1000, and 190 reviews are misclassified Using Support Vector machine, we see that 824 reviews are correctly classified among 1000, and 176 reviews are misclassified. The Comparison between Support Vector Machine, Random Forest and Random Forest Support Vector Machine algorithms (RFSVM) studied and it is very suitable for generating rules in classification methods. From the experimental results, concluded that Random Forest Support Vector Machine algorithm seems effective than the other algorithms for product reviews dataset offered by an Amazon [12].

### 3 Proposed Research Flow

In this data collection step we will collect user's reviews on various trends using social media tool. To enhancing the quality of collected data, preprocessing techniques has to be in consideration. Feature Selection as per the sentiment identification outcome also finds the features in the form of Positive, Negative and Neutral. In the data classification data will be classified as per feature selection technique and classifiers such as machine learning. After data classification data will be analyzed in the form of positive, negative and neutral. Finally on basis of Sentiment classification data will be visualized with various visualization techniques (Fig. 2 and Table 1).

In our work, proposed system will be displayed the result as per shown in Table 1. Mostly with the help of some good visualization techniques and some famous tools, system can be capable to display the final result to the user also it will help to take decision. We have to studied different dataset and it will help to check our system as per different domain (Fig. 3).

### 4 Conclusion

In an industry 4.0, Social media creates large amount of data and this data will be used for effective decision making. There is a need to many agencies and users to analyze public opinions on real time trends such as health care, social media, entertainment, product reviews, and user's interest etc., public views on social media are very important to predict future decisions. The proposed system will be used for identifying the predictive trends for decision making. In this study, the various sources of data and collection techniques are studied, The collected data needs to preprocessed for data quality



Fig. 2. Proposed Research flow

Table 1. Example of review rating summary

Aspect/Features	Extracted phrase/Text	Rating/Score	Sentiment
Movie/Web Series	Very nice, Good, Entertaining	7.8	Positive
Tourist place	Not beautiful place	2.8	Negative
Products	Fully satisfied, very good, nice	6.2	Positive
Covid Vaccination	Not effective, Risky, Good, Effective	5.0	Neutral

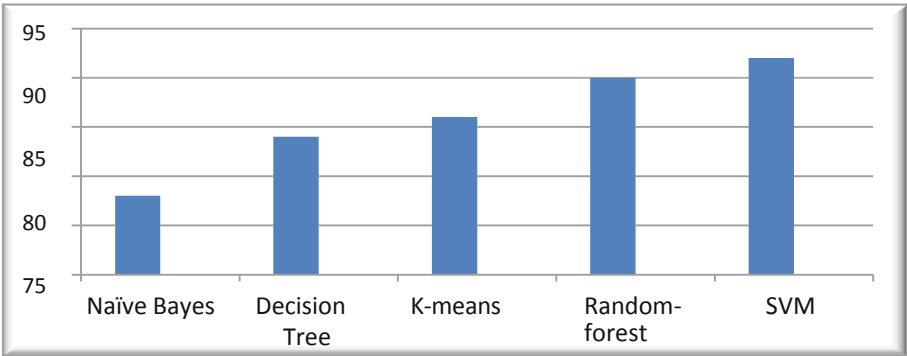


Fig. 3. Analysis on Average Accuracy in (%) percentage of Various Machine Learning Algorithm using different datasets

enhancement using various preprocessing methods. Features selection is also important to data classification. The various data classification techniques and machine learning algorithms were studied that are Naïve bayes, SVM, decision tree, KNN, Maximum entropy, Regression. Our Analysis from this review is that machine learning algorithm is provide good accuracy for various different datasets and will be useful for our proposed system also it depends on the datasets and suitable presentation of machine learning algorithm which need to be trained using collection of live data and concern is related to application domains for effective decision making. The results are represented in various forms by using effective visualization techniques as per user requirement.

## References

1. Sopan Ganpat Sutar, Intelligent Data Mining Technique of Social Media for Improving Health Care, IEEE International Conference on Intelligent Computing and Control Systems ICI CCS 2017 978-1-5386-2745-7/17. IEEE (2017).
2. A Taiping Lai#1, Xianghan Zheng Machine Learning Based Social Media Recommendation, 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM) Fuzhou, China (2015).
3. Iqbal H. Sarker, Machine Learning: Algorithms, Real - World Applications and Research Directions, Springer Nature Singapore Pte Ltd SN Computer Science (2021)
4. MALIK KHIZAR HAYAT, Towards Deep Learning Prospects: Insights for Social Media Analytics, IEEE Access (Volume: 7) INSPEC Accession Number: 18576220, Electronic ISSN: 2169-3536, (2019)
5. Donia Gamal, A Comparative Study on Opinion Mining Algorithms of Social Media Statuses The 8th IEEE International Conference on Intelligent Computing and Information Systems ICICIS ( 2017).
6. Tejas Mankar, Stock Market Prediction based on Social Sentiments using Machine Learning, 2018 IEEE International Conference on Smart City and Emerging Technology (ICSCET) Mumbai, India (2018).
7. T. Sreenivasulu Intelligent Deep Neural Network integrated with Chaotic Particle Swarm Intelligence based Sentiment Analysis in Big Data Paradigm, 7th International Conference on Advanced Computing & Communication Systems (ICACCS) (2021).
8. Dennis Thom, Can Twitter Really Save Your Life? A Case Study of Visual Social Media Analytics for Situation Awareness, IEEE Pacific Visualization Symposium 14-17 April, Hangzhou, China (2015).
9. Yuanyuan Chen, Detecting Traffic Information From Social Media Texts With Deep Learning Approaches IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 20, NO. (2019)
10. Prakash P. Rokade, Business intelligence analytics using sentiment analysis-a survey International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 1, p. 613-620 ISSN: 2088-8708, 10.11591 (2019).
11. Maria Karanasou, Scalable and Real-time Sentiment Analysis of Twitter Data 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) ISSN: 2375-9259 12-15 Dec. (2016)
12. Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri, Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis, The First International Conference on Intelligent Computing in Data Sciences, Science Direct, Procedia Computer Science 127 511-520 (2018).

13. D. Devikanniga, A. Ramu, and A. Haldorai, "Efficient Diagnosis of Liver Disease using Support Vector Machine Optimized with Crows Search Algorithm," *EAI Endorsed Transactions on Energy Web*, p. 164177. <https://doi.org/10.4108/eai.13-7-2018.164177>. (2018)
14. Santanu Mandal, Sumit Gupta "A Novel Dictionary-Based Classification Algorithm for Opinion Mining" Second International Conference on Research in Computational Intelligence and communication networks, Kolkata, India Electronic ISBN: 978-1-5090-1047-9 CD-ROM ISBN: 978-1-5090-1045-5 USB ISBN: 978-1-5090-1046-2 Print on Demand (PoD) ISBN: 978-1-5090-1048-6. <https://doi.org/10.1109/ICRCICN.2016.7813652> Page no- 175 to 180 (2016).
15. Xiaomei, Zou, et.al "Microblog sentiment analysis using social and topic context", *PLOS ONE* <https://doi.org/10.1371/journal.pone.0191163> (2017)
16. Jurgita Kapočytė, Robertas Damaševičius and Marcin Wozniak "Sentiment Analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches" *Computers*, 8, 4; <https://doi.org/10.3390/computers8010004> [www.mdpi.com/journal/computers](http://www.mdpi.com/journal/computers) (2019).
17. Xiaodong Liu , Pengcheng He , Weizhu Chen , Jianfeng Gao "Multi-Task Deep Neural Networks for Natural Language Understanding", *arXiv:1901.11504v2 [cs.CL]* (2019).
18. Snehal L. Rathod, Sachin N. Deshmukh "Sentiment Analysis Using SVM and Maximum Entropy" *International Research Journal of Engineering and Technology (IRJET)* e-ISSN: 2395 -0056 Volume: 03 Issue: 08 | page no-453-458 (2016)
19. Bhagyashri Wagh, J. V. Shinde, N. R. Wankhade "Sentimental Analysis on Twitter Data using Naive Bayes" *International Journal of Advanced Research in Computer and Communication Engineering* ISSN (Online) 2278–1021 Vol. 5, Issue 12, page no-316–319 (2016).
20. Durgesh Patel, Sakshi Saxena, Toran Verma "Sentiment Analysis using Maximum Entropy Algorithm in Big Data" *International Journal of Innovative Research in Science, Engineering and Technology* ISSN (Online): 2319–8753 Vol. 5, Issue 5, page no-8355–8361 (2016).
21. Ankur Goel, Jyoti Gautam, Suresh Kumar "Real-Time Sentiment Analysis of Tweets Using Naive Bayes" 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14–16 Electronic ISBN: 978-1-5090-3257-0 <https://doi.org/10.1109/NGCT.2016.7877424> page no-257–261 (2016).
22. Mohit Mertiya Ashima Singh "Combining Naive Bayes and Adjective Analysis for sentiment Detection on Twitter" *International Conference on Inventive Computation Technologies (ICICT)* Coimbatore, India Electronic ISBN: 978-1-5090-1285-5 <https://doi.org/10.1109/INVENTIVE.2016.7824847> page no-1–6 (2016).
23. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya "Preprocessing Techniques for Text Mining - An Overview" *International Journal of Computer Science & Communication Networks* Vol 5(1) ISSN: 2249–5789 page no-7–16 (2016).
24. I. Hemalatha, Dr. G. P. Saradhi Varma, Dr. A. Govardhan "Preprocessing the Informal Text for efficient Sentiment Analysis" *International Journal of Emerging Trends & Technology in Computer Science (IJETTCSS)* ISSN 2278–6856 Volume 1, Issue 2, page no-58–61 (2012).
25. Huma Parveen, Prof. Shikha Pandey "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm" 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) Electronic ISBN: 978-1-5090-2399-8 <https://doi.org/10.1109/ICATCCCT.2016.7912034> Bangalore, India page no-416–419 (2016).
26. Rabia Batool, Asad Masood Khattak, Jahanzeb Maqbool and Sungyoung Lee "Precise Tweet Classification and Sentiment Analysis" *IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)* Niigata, Japan, Electronic ISBN: 978-1-4799-0174-6. <https://doi.org/10.1109/ICIS.2013.6607883> page no-461–466 (2013).
27. P. Bavithra Matharasi, Dr. A. Senthilrajan "Sentiment Analysis of Twitter Data using Naive Bayes with Unigram Approach" *International Journal of Scientific and Research Publications* ISSN 2250–3153 Volume 7, page no-337–341 (2017).



28. M. Vadivukarassi, N. Puviarasan and P. Aruna “Sentimental Analysis of Tweets Using Naive Bayes Algorithm” World Applied Sciences Journal ISSN 1818–4952 Volume 35 (1) <https://doi.org/10.5829/idosi.wasj.2017.54.59> page no-54–59 (2017).
29. <https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/>

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

