Sentiment Analysis: A Simulation Based Study of Supervised Learning Algorithms.

*Himanshu Thakur, **Aman Kumar Sharma

*Department of Computer Science, Himachal Pradesh University, India.

Abstract:
Sentiment analysis is a subpart of natural language processing.Opinion mining is an effective area of research. The objective is to understand the polarity of a sentence, document to analyze the text data collected from the various sources. This paper gives an equivalent analysis of three supervised machine learning procedures (support vector machine, k-Nearest Neighbour and Random Forest) pre-owned for opinion mining on the bases of various achievement parameters. In this simulation study, it is figured out that Random Forest (RF) acquire leading fulfillment than the other two supervised learning algorithms.

Keywords: Sentiment Analysis, Opinion Mining, Supervised Machine Learning, Support Vector Machine, K-Nearest Neighbour, Random Forest, Jupyter Notebook.

Introduction

Opinion mining (OM), is also called sentiment analysis (SA),is a process of computationally determine and classify sentiments from a piece of text, and determine whether attitude towards the particular topic is positive, negative or neutral. Opinion mining is a process of analyzing unstructured text to extract relevant information. The justification OM or SA are equivalent and considered similar significance. After all, a few investigators established that OM and SA acquire a kind of various representations [1]. Sentiment analysis establishes the thought suggests in a text then consider it, while Opinion mining obtains and resolve people's assumptions about existence.The objectives of *Document-level* to allocate a response detail to communicate a positive or negative impression.*Sentence level*objective to allocate ideas communicate in an individual sentence. Essentially the first step is to determine even if the sentence is biased or unbiased.Wilsonet al. [2] obtain that learning statement is not fundamentally biased. After all inside no symbolic characteristics between document and sentence level arrangement because sentences are equitable brief documents.*Aspect level* objective to allocate opinions concerning specialized aspects of existence. Firstly it classifies the substance and then the aspects because judgment titleholder can allow the various sentiments for different aspects of the equivalent entity.The data sets pre-owned in opinion mining are necessary terminology in the present enclosure. The main originators of measurements are from the survey of production. The particular measurements are must to hold outcome according to study results of customer attitude about their production. The report origin is generally survey layout. For example, gathering consumer feedback on a marketing movement an organization can measure the movement'ssuccess or learn how to adjust it for greater success. Product evaluation is also helpful in framework better products, which can have a direct impact on the stock, as well as comparing oppositioncontributions. Opinion mining is not only enforced on a production survey but can also be activated on traditional business [4, 5], broadcastthings, [6] or officialarguments

[7].The communicative websites and online journals are expressing a brilliant authority of knowledge because of community share and exchange views on sentiments around certain matter intentionally.

In sentimentality allocation, the ultimate periodically used appearances are:

Term presence and frequency: The present appearances consist of uni-gram or n-Grams as well as their frequency.

Parts of speech Information: Identification words to the POS identity support in express differently and assist to lead feature selection.Including POS identification, the whole word in order is suggested as labels and label implement thearrangement of the word in the grammatical text.*Negation* argument performs an essential appearance to appropriate into functions because negations accept possible to reversal the opinion.*Opinion Words and phrases* are arguments and formulate which considered either positive or negative attitude. Dictionary-based and numerical based isthe fundamental approach to classify the linguistic direction of sentiment conversation.

Process of Sentiment Analysis

Sentiment analysis is a sophisticated transaction which has five phases for analysis of sentimentdata.

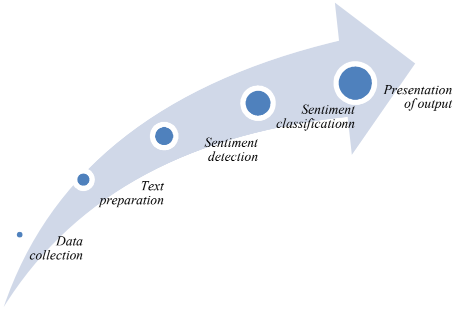

Fig.2 Sentiment Analysis Process.

Data collection: This is the first stage of SA whichcontains the collection of data from user-generated pleasure present in blogs, conference or social networks. The data is not standardized, transfer in various behaviors through the usage of various vocabularies, slang or contexts.

Text preparation: This includes cleaning of the extracted data forward to analysis. Non-textual as well as non-relevant content are identified and discarded.

Sentiment detection: Extracted sentences are analyzed and particular with a subjective opinion are retained while the resting is eliminated. Sentiment detection is done at different levels either single term, phrases, complete sentences or document with commonly used techniques such as

- Unigrams
- N-grams
- Negation
- Opinion words

Sentiment classification: Subjective sentences may be drifting out through the usage of several points.

Presentation of output: The primary aim of SA is the modification of non-structured data into useful information. When the analysis is over, text results are portrayed on graphs such as bar charts and even line graphs.

Sentiment Classification Approaches:

Sentiment allocation procedure can be branched into machine learning techniques, dictionary-based technique and hybrid technique [8]. Classification of text is done at distinct levels such as sentencebased,document-based and aspect or sentiment-based [9]. In a machine learning-based classification feel the necessity for two sets: One is Training set and another one is Test set. A training set used to set aprogrammed classifier for learning the different features of a document. Atest set is utilized for validating the performance of the classifier. The former are handled by automated classifiers for learning the differentiable features of documents while the latter is utilized for validating the performance of the automated classifier.

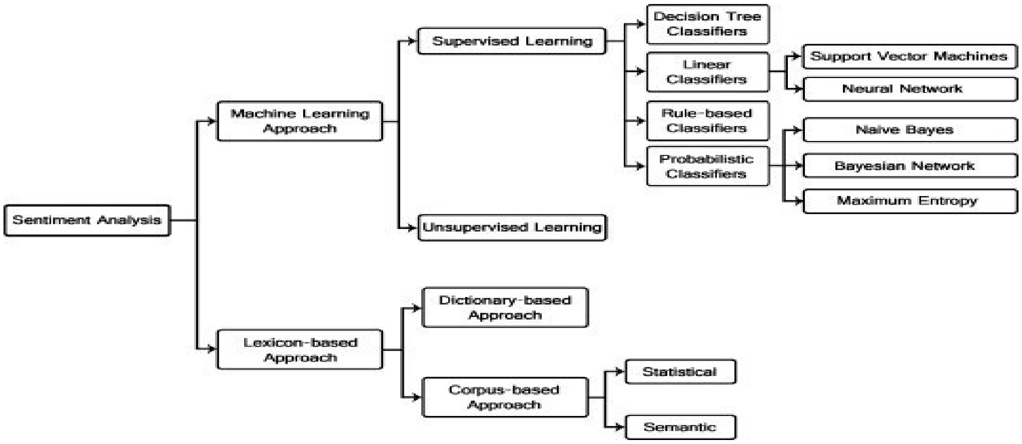

Fig.3 Sentiment Classification Techniques.[10]

(A)Lexicon Based Approach:This procedure relies on underlying sentiments (or opinion) lexicons. A sentiment lexicon is a list of lexical features (e.g., words) which are generally characterize allow to their semantic orientations as either positive or negative. Manually design andapprove such lists of opinion beingfeatures, while being among the most robust methods and alsoone of the most time consuming and available feelings expressions and management that together are called sentiment lexicon. On the other side lexicon-based technique used two methods namely the dictionary-based approach and corpus-based approach. Dictionary-based approach, a dictionary is created manually in the text editor with some necessary words, these words are collected inside the dictionary are also called seed words and they can nominate positive or negative values. After that anonyms and synonyms for these words are catch out online applyingWordNet. After the process is finished, manual analysis can be drifted out to eliminate or improve errors.The corpus-based execution advice to figure out the problem of finding sentiment words with textspecific orientations. It builds upon syntactic patterns that appear together onward with a seed list of opinions words to find other opinions words in the huge corpus. The corpus-based path is performed using the analytical or linguistic approach.

(B) Machine learning approach: Normally it depends upon the importantmachine learningmethods to figure out the SA as anapproved text distribution problem that makes use of staticor patternappearance.*Text distribution Problem Definition:* We have a set of training documents D = {Y_1, Y_2, …, Y_n} where the individualfile is classifiedinto a class.In a machine learning based classification require two sets: One is Training set and another one is Test set. A training set used to set an automated classifier for learning the different features of a document. A test set is utilized for validating the performance of the classifier. The former is utilized by automated classifiers for learning the differentiable features of documents while the latter is utilized for validating the performance of the automated classifier. Machine learning facilitatesdata processors to increase, customize and determine by itselfalthough they are identifiedto a unique input[11].The goal of machine learning is the development of compact for optimizing the performance of systems through usage of sample data or previous experiences. Machine learning attempts a solution to the classification issue which has two stages:

1) Learn the model from a training dataset.
2) Classify the test data based on the trained model. universally, classification tasks are split into various subtasks;

- Data pre-processing
- Features selection /Reduction
- Representation
- Classification
- Post-processing**.**

(1) Supervised learning: It uses a classifier; supervised learning methods depend on the presence of labeled training documents. The trained labeled documents contain words related to the topic as key features.

In supervised learning, the instruction set (An, Bn) and an innovation train a model that is valuable to conclude the achievement for each one new input. Supervised learning handles classification and backsliding performance to establish a predictive model.

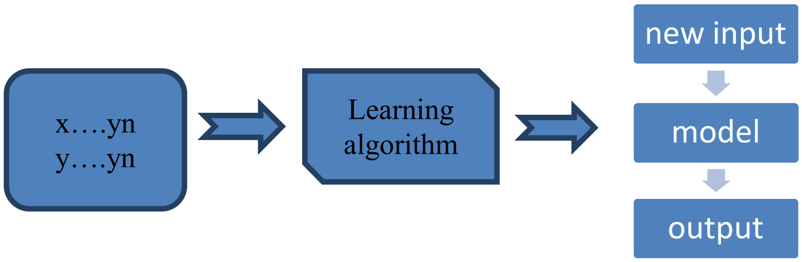

Fig.4 supervised machine learning [12]

Numerous supervised learning performancesthatacquire correlated in this paper are characterized below:A linear classifier is one that classifies something into particular classes to whatever they belong according to their appearances orattributes. A linear classifier adopts anarrow arrangement of X and Y to generate thedistribution results.

Subclasses of supervised learning include SVM(Support vector machine),K-NN(K Nearest Neighbours),and RF(Random Forest):

i. **SVM:** Support vector machine is an effective and powerful method. The SVM innovation is pre-owned as classification or regression obstacles. SVM is currently among the best performers for several classification tasks ranging from text to genomic data. SVM linearly separable binary sets. The goal is to design a hyperplane that classifies all training vectors in two classes. They are incredibly sensitive to noise in the data. SVM can be applied to complex data types beyond feature vectors (e.g. graphs, sequences, and relational data) by designing kernel functions for such data.

ii. **K-NN:** K-nearest neighbor is one of the simplest machine learning algorithms established on supervised learning techniques and mostly used for classification problems. It concludes the relationship between the new case/ data available folder and put the new case into the class that is most related to the possible categories. A K-NN algorithm stores all the applicable data and classifies a new data point established on similarity. This means when new data appears then it can be easily classified into a good suite category by using K-NN.The K Nearest Neighbours Algorithm can feel the necessity forthe division of recall to inventorygroup of the knowledge, but only put awayprediction (or study) when forecasting is essential.

iii. **Random Forest:**Random forest algorithm is a supervisedclassification algorithm. Random Forest is amplification over bagging. It uses bagging and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by the council is more accurate than that of any individual tree. Each classifier in the ensemble is a decision tree classifier and is resolvedusing a random selection of attributes at each node to resolve the split. During classification, each tree votes and the most suitable class is returned. Basically, it occupies a few drawbacks of the decision tree method, if the accuracy of the conclusion decline at the same time number of selections in tree increases.Collection of specific trees are classified to individual tree acknowledge as CART representatives (Classification and Regression Trees)

(2) ***Unsupervised Learning:*** This method is mainly used in the creation or design of trained class labeled documents thatareexamined to the most complex method. This technique is used for the document-based clustering analysis as it does not depend on pre-defined labeled training documents. The summary to be noted is that the supervised learning learns by examples whereas the unsupervised learning learns by the observation method [13, 14] inother words the unsupervisedmodel, thelearner does not get any information with solutions but in supervised learning models, examples are given to the model during the phase training.Clustering is the mechanism of combination similar entitiestogether and also used in the classification text. K-means clustering is an unsupervised machine learning algorithm. The goal of this unsupervised machine learning procedure is to find the correlation in the data point and group corresponding data points together.

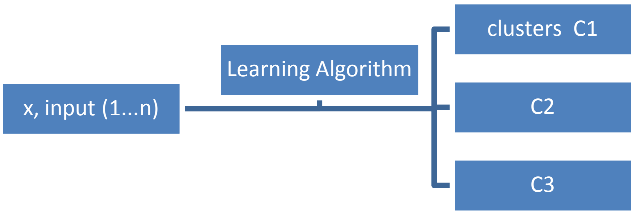

Fig.5 Unsupervised machine learning [15]

Methodology: The indicated paper examines three supervised machine learning classifiers especially SVM (support vector machine), K-NN (K-nearest neighbor) and RF (random forest) on the bases of six parameters (training accuracy, testing accuracy, total accuracy, total accuracy, F-1 score, recall, and precision). In this consideration,the Jupyter Notebook tool is used to examine these techniques.

Stepping procedure for identification based on the above-mentioned performance is inclined in fig.6:

- The name of the dataset COVID-19 (Coronavirus) is used in the study. The COVID-19 was accessed from the website kaagle (ML). [16]
- **The Jupyter** tool was used for the simulation. Jupyter is a part of Anaconda (python prepackaged distribution).
- Data preprocessing (clean and alter) the raw data in a useful and efficient format. Now use feature selections to enables the machine learning algorithms to train faster. It reduces the complexity of a model and improves the accuracy of the model.
- The data set file was COVID-19. Label encoding in python is used using sklearn library. Label encoder converts the labels into the numeric form to convert it into the machine-readable form.
- The data set file was COVID-19, Converted CSV data file into Excel data file format, and excel file obtain a practice dataset to train the effective model.
- Escape the designed mechanism for numerous supervised machine learning techniques a particular we have chosen.
- Find the results considering each technique and compare those results.

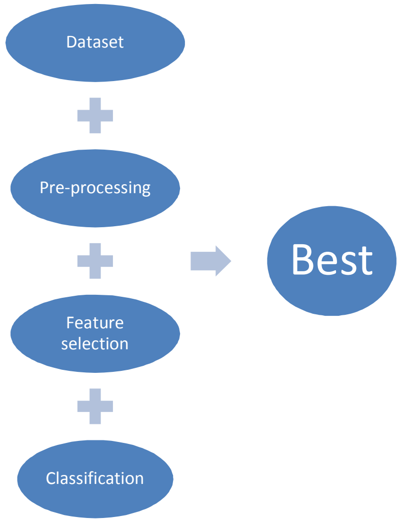

Fig.6 Task Accomplishment Methodology.

Results: The indicated classifications examine results for an equivalent study of all the three supervised machine learning techniques that are compared in this paper considering various training datasets. Entire parameters are established on *Evaluation metrics* resulted in the execution of individual classifiers.these parameters are explained as:

- **Training Accuracy**: It is usually the accuracy that applies the model on the training dataset.
- **Testing accuracy:** It is the accuracy for testing the data. Normally it is sometimes useful to compare these identify overtraining.
- **Total Accuracy**: This parameter defines the total number of the classifications divided by the total number of the documents.
  Accuracy = Number of correct predictions / Total number of predictions
- **F1 Score:** This parameter defines a weighted average of precision & recall values and calculated as:
  F=2pr / (p+r), where p=precision& r=recall.
- **Recall:** It defines that how many of true positive were recalled (found).
  Recall= Total no. of correctly classified positive examples / Total no. of positives examples.
- **Precision:** It also called positive predictive value.
  Precision = Positive correctly classified / Total predictive positive

Table 1: Comparison of all the three techniques for different parameters.

Evaluation Metrics	SVM Results	KNN Results	RF Results
Training accuracy	1	0.972	1
Testing accuracy	0.713	0.959	0.959
Total accuracy	0.713114754	0.959016393	0.959016393
F1 score	0.83253589	0.92307692	0.99428571
Recall	1	1	1
Precision	0.71311475	0.85714286	0.98863636

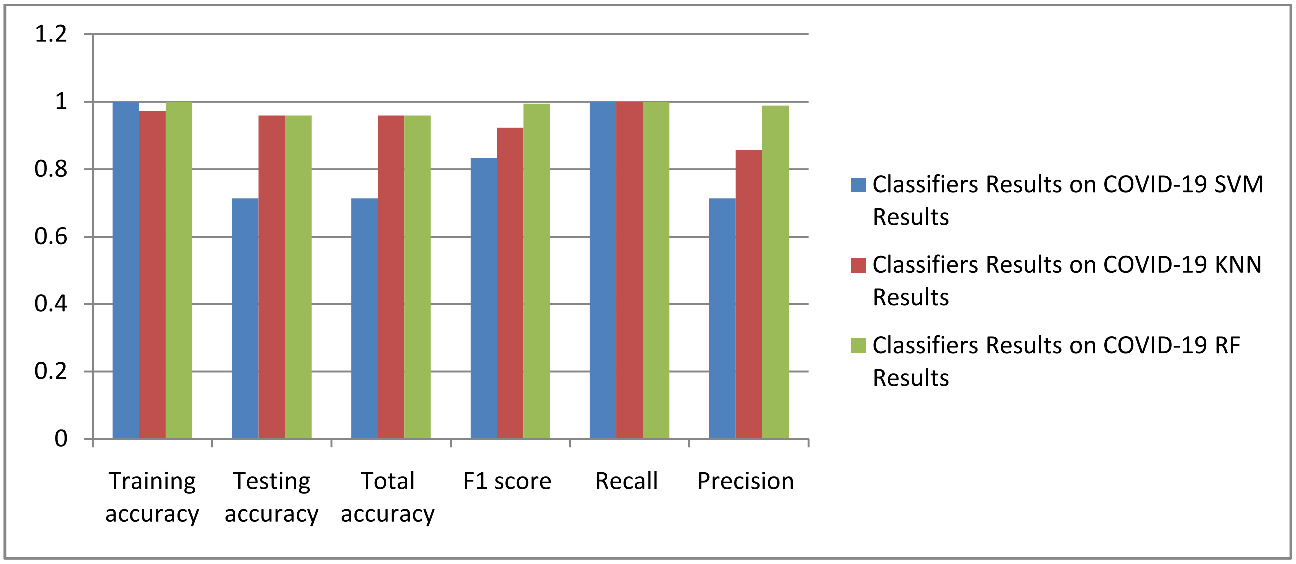

Fig.7 comparison chart

The above fig.7 shows audience outcome:

1) Random Forest (RF)acquires higher efficiency than other approaches as long as limited and wide training/sampling datasets. K-NN also has good accuracy for large datasets and SVM obtainsa regular accuracy outcome as long as classification.
2) Random Forest (RF) acquires the larger Precision meantime SVM hold fewest.
3) F1 score considering RF (Random Forest) is more superior among other approaches.

Conclusion:

Sentiment Analysis is the greater universal text classification tool that searches an incoming message and announces whether the underlying opinion is positive, negative or neutral. It iscontextual mining of textthatclassifies andselects subjective information in the source substance. This paper acquires characterized an analyzed work on Sentiment Analysis and numerous supervised machine learning algorithms. The indicated works get examined three sentiment analysis techniques: SVM, KNN and RF and include comparison them on an open-source web application tool called Jupyter Notebook. Entire these procedures are relatedto six parameters labeled as training accuracy, testing accuracy, total accuracy, precision, recall, and F1score. The presentperformance has examined coronavirus datasets thatremain used to train and verify the numerous classification models suchas individuallyhold acollection to compare.We conclude numerous results considering for entire sentiment analysis approaches and terminate that Random Forest (RF) acquire tremendous achievement than the other two techniques.

Future Scope: All the above-mentioned performances are correlated as long as the accomplishment but more future work is necessary on another developing the applications of the equivalent procedure. Existence a massive demand in commercial enterprise for the fulfillment of sentiment analysis for the reason that individual association wishes to know from what source point user vibes around the maintenance and production. In futurity performance numerous varieties of ways equivalent to supervised machine learning and dictionary-based be going to associated in regulation directed towards to defeated the interference and increase the execution through take advantage of their merits also in future; Sentiment Analysis needs extensive and wide senses corpus bases. NLP (Natural Language Processing), Google Assistant and automatic cars, etc. which put up detect the negative mood of a separate and acknowledgedpositively.

References:

[1] M. Tsytsarau and T. Palpanas, "Survey on mining subjective data on the web", *Journal of Data Mining and Knowledge Discovery*, Vol. 24, No. 3, pp. 478-514, 2012.
[2] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *In Proceedings of human language technology conference and conference on empirical methods in natural language processing,* Oct 2005.
[3] B. Lui,"Sentiment analysis and opinion mining,"*Synthesis lectures on human language technologies,* Vol. 5, No. 1, pp. 1-167, May 2012.

[4] L. C. Yu, J. L. Wu, P. C. Chang and H. S. Chu,"Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stockmarket news,"*Knowledge-Based Systems*, Vol. 41, pp. 89-97, March 2013.

[5] M. Hagenau, M. Liebmann and D. Neumann, "Automated news reading: stock price prediction based on financialnews using context-capturing features,"*DecisionSupport System*, Vol. 55, No. 3, pp. 685-697, 2013.

[6] T. Xu, Q. Peng and Y. Cheng,"Identifying the semantic orientation of terms using S-HAL for sentiment analysis,"*KnowledgeBased System*, Vol. 35, pp. 279-289, 2012.

[7] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications,"*Decision Support System*, Vol. 53, No. 4, pp. 680-688, 2012.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, 1990.

[9] A. D'Andrea, F. Ferri, P.Griffoni, and T. Guzzo,Approaches, Tools and Applications for Sentiment Analysis Implementation,"*International Journal of Computer Applications,*Vol. 125, No.3, September 2015.

[10] https://sci-hub.tw/https://www.sciencedirect.com/science/article/pii/S2090447914000550
[11] https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article

[12] https://www.ijraset.com/fileserve.php?FID=10885
[13] T. Wilson, J. Wiebe and P. Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,”*Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.
[14] B. Lui, “Sentiment Analysis and Opinion Mining,”*Synthesis lectures on human language technologies*, Vol. 5, No. 1, 2012.
[15] https://www.ijraset.com/fileserve.php?FID=10885
[16]Available [Online] https://www.kaggle.com/therealcyberlord/coronavirus-covid-19-visualization-prediction Accessed on 17/02/2020 at 12:30 PM