e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

SENTIMENT ANALYSIS OF TWITTER DATASET USING LLE AND CLASSIFICATION METHODS

Tahreem Zahra[*1], Dr Hamid Ghous[*2], Iqra Hussain[*3]

[*1,2,3]Computer Science & Engineering, Institute of Southern Punjab, Multan, Pakistan.

ABSTRACT

Twitter has become one of the most trending social media and an advantageous way to express or show emotions. Millions of Twitter users share their feelings, opinions, emotions on different aspects of life. That's why Twitter is like a gold mine of data. That data can be used for sentimental analysis, product rating, classification purpose, and wishful investigation. Different ML techniques are applied despite no other literature is on extracting features with local linear embedding (LLE). Using the twitter Application programming interface (API) with consumer and access token credentials to extract data with different hashtag classes and queries the main classes are negative and positive. In this paper, I proposed a novel hybrid framework based on feature selection method LLE and machine learning methods: Random Forest (RF), K-nearest Neighbors (KNN)and logistic Regression (LR). Random Forest (RF) has performed the best out of all.

Keywords: Locally Linear Embedding, sentiment analysis, Random Forest, K-nearest neighbor, Logistic regression.

I. INTRODUCTION

Twitter is a famous social media service which shares information in the form of tweets. Individuals express their thoughts, opinions on any substance ranging from product, individual case, etc. The process of acquiring knowledge about your views in tweet is to check your perception in general whether it's positive or negative. With the development of web technology and its evolution, there is a large amount of data available in the form of web for internet users. Many of twitter users share information, emotions and sentiments on several exposure of life.

Data mining involves method to survey information's from gigantic dataset and varies the information's in pure form. Data mining is the field of analyzing huge amount of data to discover business intelligence. data mining is the method of sorting across huge amount of data to find pattern and making relationship using data analysis. In today's companies, data mining plays a major role in forecasting business growth with the aid of analyzing data from the industry through data mining techniques.

There are two types of data mining

- Supervised
- Unsupervised

Supervised: Supervised method allow business to create model to show input and output variables. In supervised learning methods algorithm are trained using labelled data and tools used for data mining to predict future trend. There are some supervised methods:

- Support Vector Machine (SVM)
- neural network (NN)
- logistic regression (LR)
- Random forest (RF)

Unsupervised

Unsupervised learning methods create model and only input data will be given. Algorithms are not labelled data. Unsupervised learning does not use output data and less accurate.

- Cluster algorithm
- KNN
- Hierarchical

Now, we are using some methods of supervised an unsupervised like random forest, logistic regression KNN in this thesis.

we have found that none has used LLE as a feature selection method. There is a big need of using feature selection method for analysis of huge data. The feature selection can help in achieving high accuracy. That's why in this we suggest to use feature selection method LLE with Machine learning methods with twitter dataset for sentiment analysis.

Sentimental Analysis

Sentiment analysis is the zone which manages decisions, reactions just as emotions, which created through messages, being broadly utilized in fields like social media analytics, web and data mining analytics on grounds that sentiments are the most fundamental qualities to pass judgment on the human behaviour.

This specific field is making swells in both examination and mechanical social orders. Sentiments might be negative, positive, neutral or it can restrain an mathematical score clear the adequacy of the sentiment.

Sentiments might be communicated through computing the discernment of individuals on a specific point, approach and commotion to a unit [1], where this unit can be an event, a topic. Sentiment analysis and assessment mining are utilized reciprocally in a few cases however there may some events where they clasp minute disparity among them [2].

Sentiment analysis based and works on finding conclusions, characterize the attitude they pass on, and at last classify them division-wise. Surveys are first gathered all the while, its sentiment perceived, highlights chosen, sentiments ordered, and at long last sentiment polarization decided or determined.

Searching the fitting dataset is an main concern while managing sentiment analysis. It can be practical for auditing items for business, to find out the ups and lows of financial exchanges [3, 4], to get to know the mindset of individuals perusing news [5], and additionally sees communicated by individuals in political discussions [6].

Sentiment analysis is done fundamentally in light of the fact that only one out of every odd audit that is gotten offers straight "good" or a "bad" thought. Despite the fact that sentiment analysis is a lot of informative, upgrade of the examination relies upon the measure of preparing collected data that has been store into the machine.

In this paper, we discussed sentiment analysis of twitter dataset using LLE classification methods. The dataset was collected through API and processed using feature selection method LLE and classification methods. The next section, we will discuss literature review followed by methods and experiment. The last sections discuss results of those experiments.

II. LITERATURE REVIEW

In the past many researchers have discussed in sentiment Analysis of twitter data using machine learning methods. In this chapter we will review some of different research works.

Shikha Tiwari [1] described that People are used to checking something before they do it- like checking it films, pubs, shopping online and more. The present work is followed by the concept of the Twitter Sentiment Analysis, whereby the person who thinks of every particular tweet he gives is identified. They used natural language processing (NLP) has been used in current work. They found that further research work and decision making may be used. They also used Random forest and Decision tree, but they provide more precision than the SVM techniques.

Sai Ramesh et al [2] described Sentiment analysis addresses the role of analyzing natural language to decide if subjective information and the type of subjective information it express(es) are present in a piece(s) of text. The emotion behind the text reflects the subjective information: good, bad or neutral. For a large number of tweets used as big data, they have done data analysis and thus categorized the polarity of word(s), phrase(s) or whole document(s). Linear regression is used to model for the prediction of the polarity of tweeting. They work showed an 85.23% correctness. To maximize the accuracy of this method, they have applied 10-fold cross validation.

Payal Punde Et al [3] worked sentiment analysis opinion mining. They used NLP tool to sentiment analysis of twitter data. There are many aspects to the study of the sentiment analysis of Twitter info. They work on sentiment forms of research and approaches used for extracting sentiment from tweets. They found various points the study of Twitter data is performed to mine the opinion or sentiment. The principle of sentiment analysis and opinion mining was introduced in this work. This paper offers a brief understanding of tweets. WordNet semantic analysis is accompanied by machine learning techniques, such as SVM, Naïve-Bayes, the accuracy is increased. Maximum entropy, and. Using the hybrid method, precision can also be improved by up to 4-5 percent.

Dr. K. Maheshwari et al [4] worked on classification method on KNN. They work on the social network targeted is Twitter using Twitter data API. The sentiments of the tweets are classified using KNN method. The experimental results show that the classes with numerical classification produce more accurate results than text classes. According to the experiments, the accuracy of the sentiment is 36% because this attribute is text-based. The kNN algorithm classifies into 13 classes. As on other hand, the sentiment is the second-ranked attribute, therefore it produces 36%. The tuning is again needed for such work.

Dr. M. SUJITHRA et al [5] worked on Twitter sentiment analysis. They used Machine learning method using Python tool. The main aim of this work is to explore how to text analysis and determine positive, negative and neutral tweets. They improved existing sentiment analysis models further using more methods. Experimental results show that the proposed machine learning classifiers are efficient and perform better accuracy.

Ching-yu Huang et al [6] worked on twitter data referring to tweets relating to gifts, fundraising or charities. They covered strategies and methods using data analysis to capture the polarity of people's feelings about donating for any cause. They found a tweet has a neutral, negative and positive polarity by using the Natural Language Processing Toolkit (NLTK). They work as training data and use it to collect expected customers as a potential target is tweets linked to donations. They applied sentiment analysis techniques and discovered the emotion of people.

Malika Acharya et al [7] worked the famous microblog sites Twitter. The tweet(s) can be used to evaluate view(s), hence the opinion mining, extensively and efficiently. They worked sentiment analysis of twitter data using the Hadoop system is done, which analyses the vast amount of data present in the Hadoop cluster in an effective and timely manner. The main purpose of this work is to establish the actual line automate the cluster setting and then filter. Using Hadoop with Map reduce to produce excel graph. In the future, expand into more prolific language analysis such as semantical analysis, parsing, theme modelling etc. In the future.

Samantha Rai B et al [8] described the feelings of the tweets or reviews. They check for a certain tweet keyword and then compare positive and negative polarity of tweets. They used best features Naive Bayes Classifier (NBC) used to train and test word characteristics and evaluate the polarity of the feelings of each tweet. They used RF, SVM and NB classifier to be classified in a positive, neutral and nutritious way through those tweet(s) Negatives. As part of the analysis, the RF, SVM and their accuracy are taken into account. Estimated the three features and thus the number of tweets increased. In addition, RF accuracy, Extend the number of tweets is also estimated at SVM and NB.

Rajinder Singh et al [9] present the Words and phrases on social media related to many topics reflect the people's viewpoint. They find positive polarities or negative. They used NLP to explain the text posted on media platform of sentiment analysis. They used four machine learning methods. For reviewing of sentiment analysis. Optimization Naive Bayes, J48, Tree, and OneR. They are used two Datasets, Amazon and one dataset from IMDB movie reviews. They found effectiveness of four of these techniques' characterization is investigated and analyzed. In learning, the Naïve Bayes, it was found to be very quick, while OneR appears to be more promising in producing 91.3 percent precision accuracy, 97 percent F-measure accuracy, 92.34 percent instances correctly classified for accuracy. According to the experimentation outcomes. Naïve Bayes shows a quicker learning rate, true/false positive levels in the actual, J48 discloses adequacy. For smaller Woodland wallet overview datasets, J48 and OneR are better. They concluded, In Future studies of emotions (sentiment analysis) task has the potential to enhance pre-processing using deep neural networks with word embedding's, this research can also be extended by convolution neural network(s).

Dae Ryong Seo et al [10] presented subjects in the field of data handling is conclusion examination or feeling mining. It looks to choose whether the extremity of text information (report, sentence, section) will bring about a positive, negative or neutral characteristic. They used document text from twitter about the Indonesian film review. They used Bag of words, and approach used Naïve Bayes and Ensemble Features. There are numerous different sorts of highlights that have been utilized for this gathering. Ensemble Features is 0.88. In the interim, Bag of Words highlights with 0.94 f-measure esteem have better productivity. They work on naïve Bayes classifiers, film review and divided into positive and negative tweets. They found Naïve Bayes showed strong performance in this work, and found that the Bag of Word feature(s) with 0.96 accuracy,0.92 recall, and 0.94 f-measure value have the best performance among these features

In this paper, a novel hybrid framework is proposed using feature selection method LLE and machine learning methods: random forest, logistic regression and k-nearest neighbors.

III. MODELING AND ANALYSIS

Sentiment analysis is the automatic method of defining and categorizing contextual details in text data. Sentimental study is the process of assessment and of having the viewpoint or attitude of the authors. This is a view, a decision, or a perception about a particular aspect of a topic or product. The most common type of sentiment analysis is polarity detection, which classifies a statement as 'positive',' negative' or 'neutral'.

In this chapter, a novel hybrid framework is proposed using feature selection method LLE and machine learning methods: random forest, logistic regression and k-nearest neighbors.

Dataset

A Dataset is a collection or collection of information. Normally, this set is presented in a tabular pattern. A specific variable is defined in every column. And each row, as per the given query, corresponds to a given member of the data set. In this segment, we collect data about Twitter using the Twitter API through the Developer account. Total dataset are 31962 tweets in training datasets in which the positive values are “29959” and negative values are “2003”. When we collected data, it consisted of four confusion class e.g happy, sad, cheery and hatred after organizing the data we made a conclusion that the data must be in binary form. For which we made happy and cheery as “Positive”, sad and hatred as “Negative”. This is a part of data management.

We get the data using the following steps.

1) Credential
2) Consumer Key
3) Consumer Token
4) Access Token

Training dataset:

For the preparation of our dataset, we implemented this technique on 31962 attributes. In order to serve as a basis for further implementation and utilization, neural networks and other artificial intelligence systems need an initial collection of data, called training data. This data is the cornerstone of the growing library of knowledge for the programmed. To train these models we used Logistic Regression, KNN and Random Forest techniques. When a model on a training set is conditioned, it is normally tested on a test set. These sets are always taken from the same overall dataset, but the training set can be named or enriched to improve the trust and accuracy of an algorithm.

There really isn't a hard-and-fast rule about how much information you need. However, as a general rule of thumb, you would need more details than you think. Different use cases will require various quantities of data, after all. One where you need to be extremely sure about your model would require massive quantities of data, while a comparatively narrow text-based sentiment model needs much less data.

The model blends into a data set for planning initially. Parameters of the algorithm the algorithm is conditioned on the training dataset using, for example, optimization approaches such as gradient descent or stochastic gradient descent using a supervised learning framework. The training data set consists mainly of input (or scalar) pairs of vectors and the corresponding output (or scalar) vector in operation, where the solution key is normally referred to as the target (or label). With the training data collection, the current model is run and produces a result in the training dataset for each input vector, which is then compared with the target. The parameters of the model are modified depending on the effect of the comparison and on the same learning algorithm being used. Both variable collection and parameter estimation can be used in model fitting.

Testing dataset:

By applying research techniques on 17190 attributes, we checked our data collection. Finally, the Evaluation Data Set data set is a data set that is used to provide an impartial indicator of the final model that falls into the training data set. To test these models we used Logistic Regression, KNN, Random Forest techniques and we got the maximum accuracy in the Random Forest technique. A holdout data set is often known as a test data set if the data in the test data set (for example, in cross-validation) has never been used in testing.

A data set for training is a data set of examples used during the learning process and is used, for example, to tune a classifier's parameters. Most methods that look for observational similarities through training data appear to over-fit the data, meaning that it is possible to find and manipulate apparent connexons that do not actually occur in the training data.

Table-1: Few ID's are mentions above in which Label with 0 values represents positive statement and label with 1 value represents as a negative statement.

Id	Label	Tweets
1	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo
2	0	@user @user Zahra is a good girl! i'm it's so #gr8!
3	1	@user names 'create the wall' chant "# tcot" # Michigan middle school
4	1	@user @user lumpy says i proved it, lumpy.

Pre-Processing:

A significant step in the process of data mining is data pre-processing. For data mining and machine learning ventures, the term 'garbage in, garbage out' is especially applicable. Data collection methods are also poorly regulated, leading to unlikely data combinations and missing values, etc., resulting in out-of-range values. Analysing information that has not been carefully screened for such issues will yield misleading results. Therefore, prior to running any research, the representation and consistency of data is first and foremost. The most important step of a machine learning project is often data pre-processing, especially in computational biology.

If there is a lot of unnecessary and redundant information available or noisy and inaccurate information, so it is more difficult to find expertise during the training process. It can take considerable processing time for data preparation and filtering steps. Pre-processing of data involves washing, collection of instances, normalization, transformation, extraction of features and set Tasks of Pre-processing Data

- Acquiring the data set
- Importing all the essential libraries
- Importing the data set

Feature Extraction

Function extraction is the term for strategies that select and combine variables into functions, thus reducing the amount of data to be processed, but also accurately and completely representing the original data collection. The extraction of qualities starts from an underlying assortment of determined information in AI, design acknowledgment and picture handling, and delivers inferred values (highlights) planned to be expressive and non-excess, encouraging resulting proportions of learning and clearing speculation and adding to better human translations in specific cases. Extraction of capacities is because of dimensional decrease. The chose highlights will contain the necessary info information data, so the ideal errand can be cultivated by utilizing this diminished portrayal rather than the full introductory information. We apply three basic techniques on the given dataset which are shortly described below.

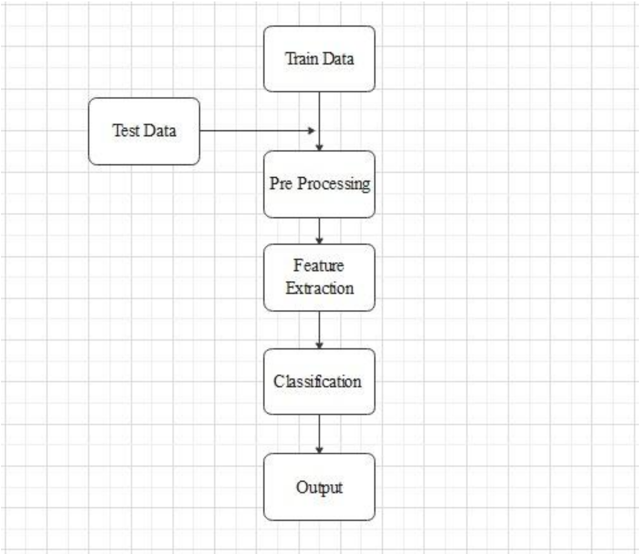


Fig.-1: The diagram shows the methodology of the work

Random Forest or Random Decision Forest

Random forest or random decision forests is learning method of classifying, regressions or the different things that deal by building host of decisions tree during the time to train and output the classes that are the form of the class or standard predictions in the individual tree. For decision trees' habit of fitting to their preparation, random decision forest is precise. In the decision trees, Random forests typically perform very well, but their perfection is very poor than the gradient boosted trees. Data characters, however, can affect precision. The techniques for supervised machine learning techniques are these algorithms. The popular techniques for different machine learning tasks are the Decision Tree. Tree learning comes closest to meeting the requirements for dealing with data mining as an off-the-shelf method.

In particular, deep trees appear to get knowledge of highly erratic patterns: they suit their patterns of training, i.e., they have less bais but rather high difference. Random forest is way to average several deep decision trees, which are trained with the motto of minimising variation on non-identical sections of the identical training set [12]. This comes at the expense of a slight rise in bais and no lack of transparency, but in the final model, it typically improves the outcomes. Forest is like the Decision Tree algorithm effort moving together. Moving the teamwork of several trees will enhance a solitary random tree 's output. Although not identical, forests offer the K-fold cross reliability effect. These following steps are used in the random forest techniques of the machine learning: -

1. Firstly, we are going to begin by selecting random sample from any of the given dataset.
2. Next, for each sample, these algorithms will create a decision tree. Then the prediction outcome is derived from every decision tree.
3. For every predicted outcome, voting would be carried out in this phase. Finally, pick the outcome of the most voted forecast as the final outcome of the forecast [12].

Logistic Regression

In mathematics, the Logistic Regression Model or Logit Regression is used to model the possibility of a certain class of event circumstance, such as passing of falling, winning or losing, alive or dead, healthy or bad [13]. This can also be used to model different kinds of activities, such as identifying when a cat, dog, lion, etc. has an image. The logistic regression model is mathematical in its early form, and the model means a logistic equation to the binary dependent variable model, there is also a more complicated expansion. Logistic regression (or logit regression is an approximate logistic model parameter that is a binary regression form) in regression analysis. A binary logistics model has a consistent variable represented by the same variable with two possible values, such as pass or fail, where '0' and '1' are labelled as two values.

- **Binomial or binary logistics** regression deal with conditions where there can only be two possible forms of the observed effect of a dependent variable," 0 "and" 1(which represents, e.g, "alive" or "dead" and "loss" or "win").
- **Multinomial logistics regression** deal with circumstances in which only three or more potential forms of outcomes can be achieved (for example., "stage A" vs. "stage B" vs. "stage C") those are never ordered.
- **Ordinal logistic regression** deal with the conditional variables those are always ordered [13].

k-Nearest Neighbors Algorithm (KNN)

The K-nearest Neighbors (K-NN) algorithm is non-parametric approach used by Thomas Cover for classifications and regressions [14]. The inputs, in either case, consist of the nearest example of k training in the function spaces. Output depends on the weather used for k-NNN regression or classification:

- Class membership in k-NN classifications is the performance. An object should be ranked by the plurality of votes of its neighbor, with objects being distributed among the k of its closest neighbor to the most common class (k is a positive integer, maybe little). For k = 1, the object is given to singular nearest - neighbor class in a simplified way.
- In k-NN regressions, the outputs are the properties and values of the objects. These values are the averages of values from k nearest neighbor(s).

K-NN is a form of example-based learning, or slow learning, where functions are only estimated locally and all computations are delayed prior to the functions' assessment(s). Because this type of algorithm depends on classification distances, the normalization of all data training would increase its dramatic accuracy.

The Euclidean distance for dependent variable is a commonly used distance metric. For discrete variables, such as for text categorization, other metric may be used, including the overlapping metric (or Hamming distance). For example, in terms of genomic data sets with correlations, such as Pearson and Spearman, k-NN has been used as a parameter. It is also possible to effectively increase the classification accuracy of k-NN if the distance approximation is studied with complex algorithms such as Broad Margin Nearest Neighbor(s).

- In k-NN regressions, the outputs are the properties and values of the objects. These values are the averages of values from k nearest neighbor(s).
- Class membership in k-NN classifications is the performance. An object should be ranked by the plurality of votes of its neighbor.
- The objects being distributed among the k of its closest neighbor to the most common class (k is a positive integer, maybe little). For k = 1, the object is given to singular nearest - neighbor class in a simplified way [14].

Locally Linear Embedding (LLE)

This is an unsupervised learning algorithm, which calculates low-dimensional neighborhood preservation of high-dimensional data embedding, LLE is described here In high-dimensional data, LLE tries to leverage the local symmetries of linear reconstructions, to find a nonlinear structure [16]

Below given is the diagram that shows the complete working of the system including the models and LLE:

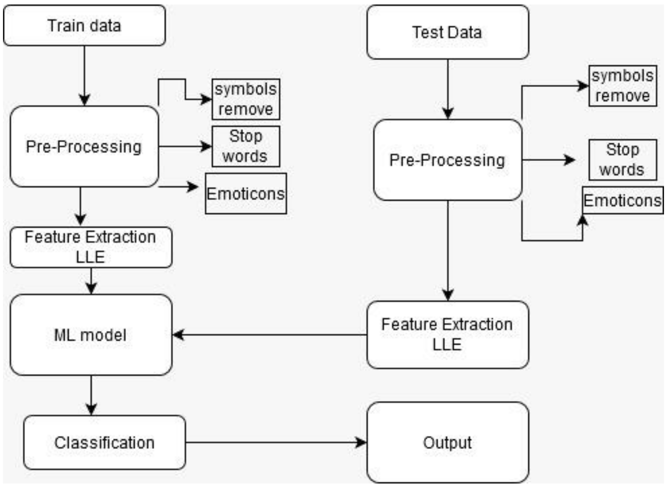


Fig.-2: Model Representation

IV. RESULTS

We are using ANACONDA (2017). With Python 3.7, Anaconda 5.3 is compiled, taking full advantage of the speed and feature enhancements of Python. In the latest release, Anaconda 's reliability has been improved by capturing and storing item information for installed packages. Here for twitter dataset we are using the developer account to get the data from the twitter API. Here we are implementing machine learning and in machine learning we are using three different techniques Random Forest, KNN and Logistic Regression.

Random forests or random decision forest is learning method of classifying, regressions or the different things that deal by building host of decisions tree during the time to train and output the classes that are the form of the class or standard predictions in the individual tree. Logistic Regression Model or Logit Regression is used to model the possibility of a specific class of event happening, such as passing of falling, winning or losing, alive or dead, healthy or bad. K-NN is form of example-place investigation, or slow learning, whereas functions are only estimated locally and the all computations are delayed prior to the functions' assessment(s).

Implementation of experiment

We are using anaconda (2017) that provide many tools for the classification of tweets. Here for the tweeter dataset we are using machine leaning techniques (Jupiter tool). Here we different tweets datasets by using different tools. The data is divided in to two class 0 and 1. These two sets having different datasets 0 having positive tweets and 1 having negative tweets.

Experiment No 1

The experiment is done on the twitter datasets. Two datasets are used in this experiment one is training datasets on which we have trained the machine learning models, Second is testing dataset on which we have test the trained model. On this twitter data we apply the Random Forest, KNN and Logistic Regression Techniques. On this model evaluation is performed and here we check the accuracy of twitter datasets using different models. Here the accuracy is 0.68 for KNN, 0.69 for Random Forest and 0.70 for the Logistic Regression technique.

Following Model are applied on these Twitter datasets

1. KNN.
2. Random Forest.
3. Logistic Regression

1) KNN

The K-nearest Neighbors (K-NN) algorithm is a non-paramedic approach used for classifications and regressions. The inputs, in either case, consist of the nearest example of k training in the function spaces. Output depends on the weather used for k-NNN regression or classification:

- Class membership in k-NN classifications is the performance. An object should be ranked by the plurality of votes of its neighbor, with objects being distributed among the k of its closest neighbor to the most common class (k is useful integer, maybe little). For k = 1, the thing is given to the singular nearest - neighbor class in a simplified way.
- In k-NN regressions, the outputs are the properties and values of the objects. These values are the averages of values from k nearest neighbors(s).

K-NN is a form of example-based leaning, or slow training, where functions are only estimated locally and computations are delayed prior to the functions' assessment(s). Because this type of algorithm depends on classification distances, the normalization of all data training would increase its accuracy. We get the accuracy of KNN

Table 2:

	precision	recall	f1-score
Positive	0.35	0.97	0.52
Negative	0.99	0.76	0.86

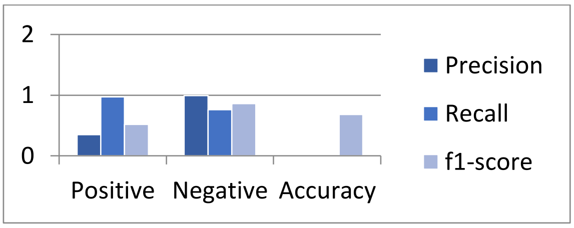


Fig.-3: Graph Representation of KNN

2)Random Forest

Random forests or random decision forest is learning method of classifying, regressions or the different things that deal by building host of decisions tree during the time to train and output the classes that are the form of the class or standard predictions in the individual tree. For decision trees' habit of fitting to their preparation, random decision forest is precise. In the decision trees, Random forest's typically perform very well, but their perfection is very poor than the gradient boosted trees. Data characters, however, can affect precision. The techniques for supervised machine learning techniques are these algorithms. The popular techniques for different machine learning tasks are the Decision Tree. Tree learning comes closest to meeting the requirements for dealing with data mining as an off-the-shelf method. We get the accuracy of Random forest is 0.69%.

Table 3

	Precision	Recall	F1-score
Positive	0.44	0.88	0.59
Negative	0.97	0.79	0.87

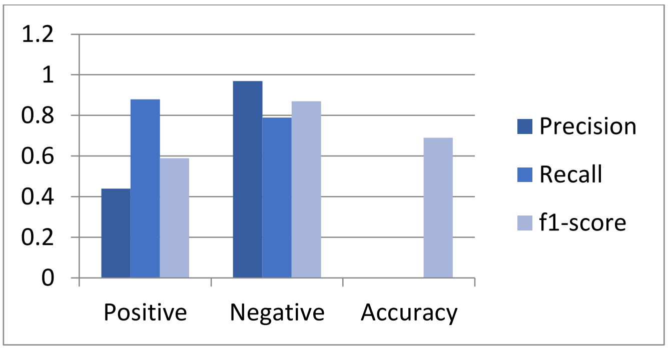


Fig.-4: Graph Representation of Random Forest

Logistic Regression

The Logistic Regression Model or Logit Regression is used to model the possibility of a certain class of event happening, such as passing of falling, winning or losing, alive or dead, healthy or bad. This can also be used to model different kinds of activities, such as identifying when a cat, dog, lion, etc. has an image. In its early form, the logistic regression model is statistical Model implies a logistic equation to model binary dependent variable, there is even a more complex expansion. In regression analysis, logistic regression (or logit regression is an estimated logistic model parameter that is a binary regression form). A binary logistics model has a consistent variable represented by the same variable with two possible values, such as pass or fail, where '0' and '1' are labelled as two values. Now, we get the accuracy of logistic Regression is 0.70%.

Table 4

	Precision	Recall	F1-score
Positive	0.24	0.87	0.38
Negative	0.98	0.75	0.85

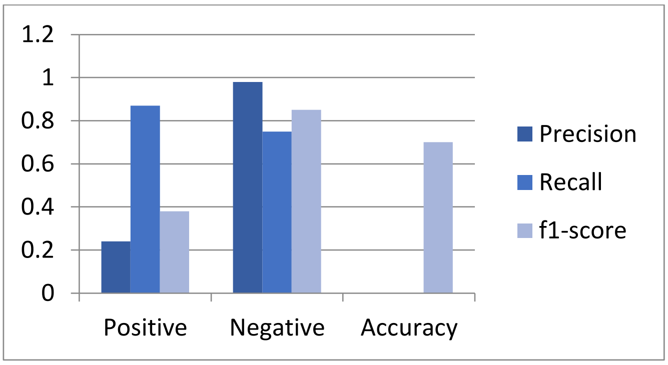


Fig.-5: Graph Representation of Logistic Regression

Experiment No 1 Result Table 5:

Algorithms	Accuracy	Time to build model
Random Forest	0.69	8 sec
KNN	0.68	9 sec
Logistic Regression	0.70	4 sec

Experiment No.2

In this experiment, we used LLE as a feature selection method before applying machine learning datasets by using the LLE technique. On this twitter data we apply the Random Forest, KNN and Logistic Regression Techniques. On these model's evaluation is performed and here we check the accuracy of twitter datasets using different models. Here the accuracy is 0.79 for KNN, 0.80 for Random Forest and 0.76 for the Logistic Regression technique using LLE which was 0.68, 0.69 and 0.70 respectively before applying the LLE Technique. We will check the data on training validation and testing base.

1) KNN

Table 6: After applying, LLE feature selection method we get the accuracy of KNN 0.79%.

	precision	recall	f1-score
Positive	0.35	0.97	0.52
Negative	0.99	0.76	0.86

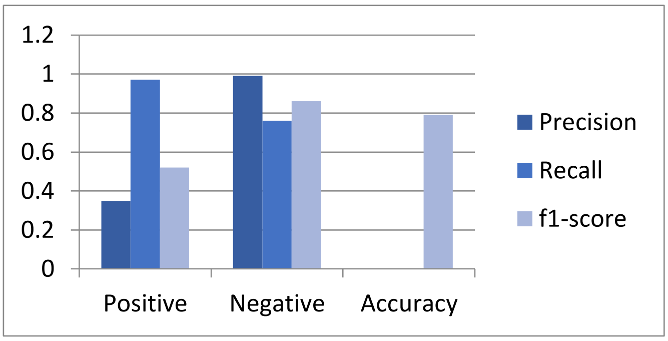
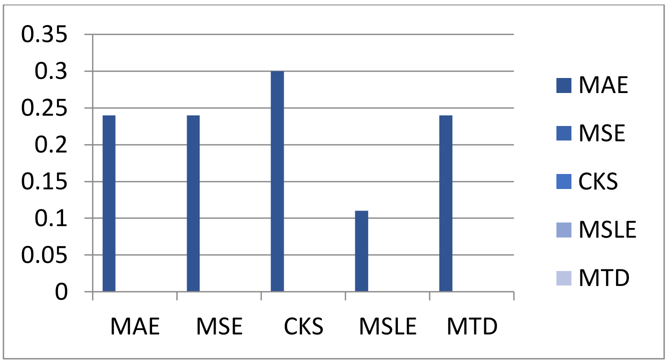


Fig.-6: Graph Representation of KNN

Confusion Matrix

Table 7

	Positive	**Negative**
Positive	63	115
Negative	2	372



2) Random Forest

Table 8: After applying, LLE feature selection method, we get the accuracy of Random forest 0.80%.

	Precision	Recall	F1-score
Positive	0.44	0.88	0.59
Negative	0.97	0.79	0.87

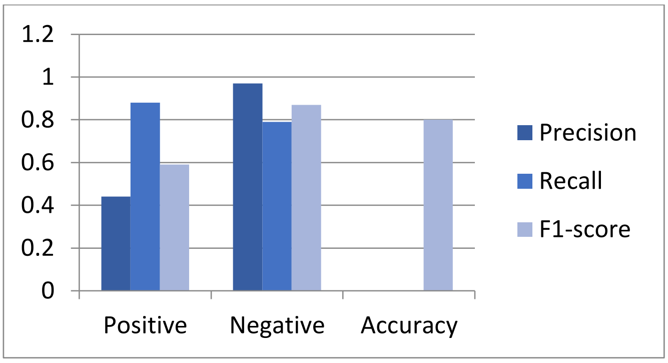
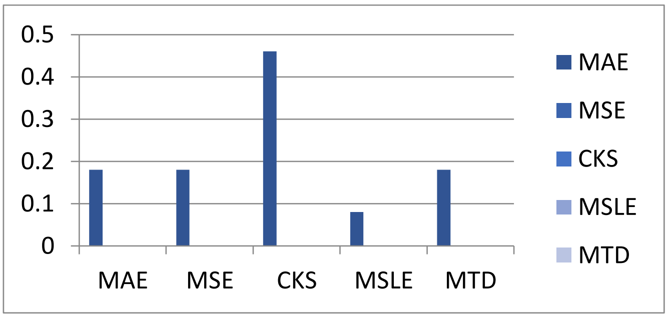


Fig.-7: Graph representation of Random Forest

Confusion Matrix

Table 9

	Positive	**Negative**
Positive	115	146
Negative	16	551



3) Logistic Regression

Table-10: After applying, LLE feature selection method. We get the accuracy of Logistic Regression 0.76.

	Precision	Recall	F1-score
Positive	0.24	0.87	0.38
Negative	0.98	0.75	0.85

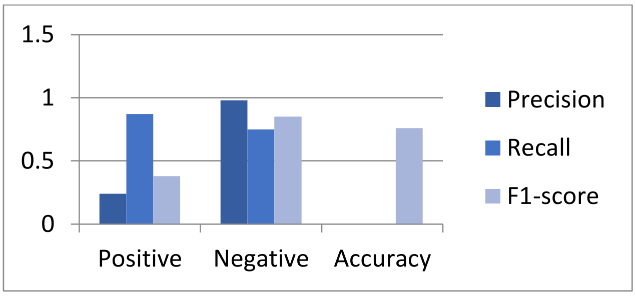
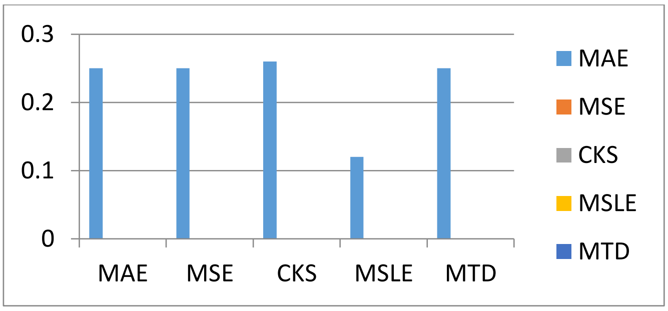


Fig.-8: Graph Representation of Logistic Regression

Confusion Matrix

Table 11

	Positive	**Negative**
Positive	61	191
Negative	9	567



Results graph after applying LLE Technique

Table 12

Algorithms	Accuracy	Time to build model
Random Forest	0.80	3 sec
KNN	0.79	6 sec
Logistic Regression	0.76	8 sec

V. CONCLUSION

Twitter is one of the most used communication platform networks in which the officials use their account to make people aware of their deeds. According to the research there are 321million active users in the world but twitter is mainly used by the officials. We are presenting a novel hybrid framework for the classification of twitter dataset which we got from the twitter API it was approximately 18968 tweets using KNN and machine learning techniques. The objective of this research is to choose the algorithms and metrics for comparing the overall performance of Machine Learning Classifiers and to examine the metrics acquired from special system getting to know algorithms relying on the dimensions of datasets.

First Experiment was done using the KNN, Random Forest and Logistic Regression technique in which we got the low accuracy but after applying the LLE technique we acquired better accuracy then before. Our data was in binary form e.g happy or cheery as “1”, sad or hated as “0”. The tools we used were ANACODA using python language. The best results was got with the Random Forest technique after using Locally Linear Embedded (LLE) method.

In future we want to try more machine learning techniques on large dataset. We also want to work on multi classes on different dataset of twitter using different machine learning techniques.

VI. REFERANCES

[1] Shikha Tiwari, AnshikaVerma, PeeyushGarg, Deepika Bansal. P., 2020. Social Media Sentiment Analysis On Twitter Datasets.2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS).

[2] A Razia Sulthana, A K Jaithunbi, L Sai Ramesh P., 2018. Sentiment analysis in twitter data using data analytic techniques for predictive modelling. National Conference on Mathematical Techniques and its

Applications (NCMTA 18) IOP Publishing. IOP Conf. Series: Journal of Physics: Conf. Series 1000 (2018) 012130 doi :10.1088/1742-6596/1000/1/012130

[3] RasikaWagh, Payal Punde.P.,2018. Survey on Sentiment Analysis using Twitter Dataset. Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1

[4] Dr. K. Maheswari. P., 2018.Improving Accuracy of Sentiment Classification Analysis in twitter Data Set Using knn.E ISSN 2348 –1269, PRINT ISSN 2349-5138

[5] S.Siddharth, R. Darsini, Dr. M. Sujithra. P., 2018. Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python. Research Gate.

[6] Amrita Shelar and Ching-yu Huang. P., 2018.Sentiment Analysis of Twitter Data.2018 International Conference on Computational Science and Computational Intelligence (CSCI).

[7] Malika Acharya, Shilpi Sharma. P., 2018.Semantic Analysis of Twitter Posts.978-1-5386-7709-4/18/$31.00 c 2018 IEEE

[8] ShamanthaRai B, Sweekriti M Shetty. P., 2019. Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance. 2019 IEEE 4th International Conference on Computer and Communication Systems

[9] Jaspreet Singh, Gurvinder Singh, and Rajinder Singh. P., 2017.Optimization of sentiment analysis using machine learning classifiers. Singh et al. Hum.Cent. Comput.Inf. Sci. (2017) 7:32 DOI 10.1186/s13673-017-0116-3.

[10] Rosy Indah Permatasari, M. Ali Fauzi, Putra PanduAdikara. P.,2018. Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes.

[11] Koyel Chakraborty, Aboul Alla Hassanien, in Social Network Analytics, 2019

[12] T. Shi and S. Horvath. Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1):118-138, March 2006.

[13] Freund R.J. Wilson W.J. (1998). Regression Analysis: Statistical Modeling of a Response Variable. San Diego, Academic Press

[14] J. Zhang and I. Mani, "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction", Proc. Int'l Conf. Machine Learning (ICML '2003) Workshop Learning from Imbalanced Data Sets, 2003.

[15] de Ridder D, Kouropteva O, Okun, Oleg (2003) Supervised locally linear embedding. In: Proceedings of ICANN, pp 333–341