IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TRADITIONAL DATA MODELING WITH NATURAL LANGUAGE PROCESSING AND EVALUATIONS

[1]K. Suguna, [2]Dr.K. Nandhini

[1]Research Scholar, [2]Assistant Professor

[1]Department of Computer Science, [2]PG & Research Department of Computer Science

[1]Bharathiar University, [2]Chikkanna Government Arts College

[1]Coimbatore- 641046, *[2]Tirupur-641602, [1]India, [2]India.*

***Abstract*- The data analysis process is a major protagonist when dealing with huge amounts of data. The analysis comprises data from various sources and manipulating them using data analysis tools and techniques. This research work concentrates on prediction learning by analyzing web data using Baye's theorems. It generates the conditional pattern base to identify the event's occurrence in each node and stores the folds of how many times that particular event has occurred. The evaluation methods predict the data in which the highest conditional pattern base has been generated. The proposed system deals with a real-time Twitter dataset (geographical data) which is enormous and more complicated to predict the interest of the user.**

Keywords: Big data, Data Mining, Natural Language Processing (NLP), Naïve Bayes Classification.

I. INTRODUCTION

The number of users of shared networks is increasing rapidly. Applying analysis methods for predicting information from massive amounts of big data is more complicated. Data Mining is the analysis of large quantities of data to extract previously unknown, interesting patterns of data, unusual data, and dependencies. By increasing the diagnostic accuracy of the classification methods, error classification can be used [1]. Intellectual classification methods consist of computer-assisted artificial intelligence-based algorithms [2]. This study aims to achieve the highest accuracy in prediction by using a combination of traditional and intelligent methods using the same data. The data stream is a combination of small and big data like personal, common, open, inaccessible, private, viable, and authorized data [3][4]. Big Data is frequently virtual to the initial stage of an organization and when its ability to handle data using existing systems becomes no longer feasible [5]. A shared database defines a data repository used for research and housing data related to technical research on an open platform. The databases collect and accumulate diverse and multi-dimensional data and perform systematic research in a structured form [6]. The specific analytic methods are to be used for dissimilar data sources with detailed structures and perform the analysis based on the general features of records [7]. Note that the goal is the extraction of patterns and knowledge from large amounts of data and not the extraction of data itself [8]. Descriptive analysis uses statistical approaches, classification methods to isolate data, estimate to predict, and various other techniques using data mining techniques. Data analysis has various applications widely applied in the multi-domain industry [9][10]. Predictive learning becomes more personal, preventive, and sharing. AI can make main contributions in these fields [11]. Causal inference is a powerful modeling tool for explanatory analysis, which might enable current machine learning to make accurate predictions [12]. The NLP model provides a more efficient system for dealing with big data [13]. The patterns obtained from data mining can be considered as a summary of the input data that can be used in further analysis or to obtain more accurate prediction results by a decision support system [14]. The documented text is then semantically parsed into logical forms that can be used to automatically extract the answer from the underlying database.[15]. A new method of variable selection was used for the Naive Bayes classifier to test it on a large set of popular datasets. The performance and experiments were developed to compare the results for Naive Bayes using a variable selection procedure with different thresholds. A new prediction method was found to give substantial results with a very true positive with high occurrence unless the sample size is very small.

II. FEATURE EXTRACTION

Feature selection is the process of extracting the important features to improve the learning process and efficient prediction by selecting the relevant text and removing irrelevant text. It enables the user to build useful models without modifying the original data. The feature selection method may be with the class labels or it may not be with a class label. The method with class labels is used to classify the texts with relevant features. Some other methods are used for text without class labels.

The method can be a filter method or wrapper method or embedded method. The properties of the filter method are measured through statistics.

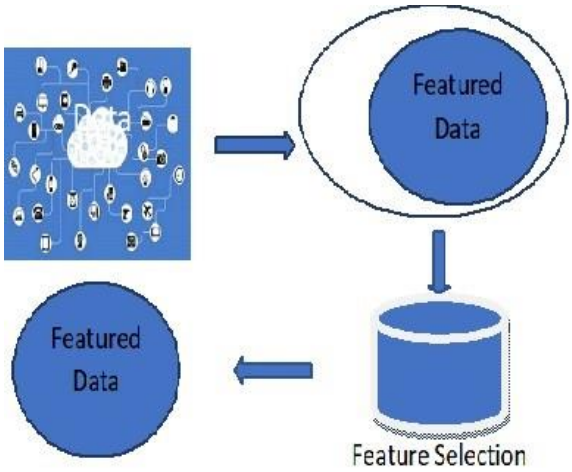

Figure 1. Feature Selection Process

It is cheaper to compare with other methods in dealing with high-dimensional data sets. The wrapper method creates a space for subsets and evaluates their quality. The embedded method is a combination of the filter and wrapper method which provides the best features at a reasonable cost.

A. Term Frequency (TF)-Inverse Document Frequency (IDF)

TF-IDF compares the number of times a term appears in a document with the number of documents the word appears in. It also predicts the possibility of finding a word within the document. The term frequency is calculated as

$$tf(t_i, d_j) = \frac{No.\,of\ times\ t_i\ occurs\ in\ d_j}{Total\ No.\,of\ words\ in\ d_j} \quad (1)$$

$$tf(t,d) = 1 + log f_{t,d} \quad (2)$$

The term frequency *tf* holds the *term t* and the number of times the *term t* appears in *document d.* The inverse document frequency contains the number of documents and the document frequency of the *term t.* The inverse document frequency is a measure of a term that occurs frequently in the entire text. *IDF* is a standard log value, attained by dividing the total number of documents in the entire dataset by the number of documents containing term *t*, and calculating the logarithm of the entire term.

B. Document frequency (DF)

Document Frequency (DF) is the count of occurrences of *term t* in the document set *N.* It contains several documents in which the word is present. The term consists of the document at least one it is considered one occurrence.

$$df(t) = occurence\ of\ t\ in\ d \quad (3)$$

C. Inverse Document Frequency (IDF)

An *inverse document frequency* feature decreases the weight of terms that occur very frequently in the document and increases the weight of terms that occur rarely. It inverses the document frequency which measures the *term t.* The relative weightage is obtained in this phase. The df will be 0 when there is no vocabulary in a document. instead, 1 will be added to the denominator.

$$idf(d, D) = \log \frac{|D|}{\{d \in D : t \in D\}} \quad (4)$$

The tf-idf is capable to measure to evaluate words in the document. The basic computation task of the tf-IDF is accomplished through

$$tfidf(t, d, D) = tf(t, d) * idf(d, D) \quad (5)$$

The TF-IDF is used in text-processing applications in Natural Language Processing (NLP). To improve the accuracy of this model the selected number of a subset of features is incorporated.

III MODEL BUILDING

The NLP model is built using the Naïve Bayes classification algorithm. The naïve Bayes classifier algorithm uses the Bayes theorem to classify the objects. This theorem predicts the strong or weak attributes of the actual data. It is a basic mathematical formula for probabilistic machine learning algorithms used in a wide range of NLP problems to find the measure of conditional probabilities.

A. Conditional Probability using Bayes Theorem

The probability of the existence of an event related to any condition is identified. It defines the probability of occurrence of any event 'A' when another event 'B' about 'A' has already occurred. The probability of A occurring when another event B has already occurred is identified as shown in (6).

The probability of 'B' occurring in the given dataset in which 'B' has already occurred.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (6)$$

It is defined as P(B|A). It will be the multiple of the probability of an 'A' occurrence. Then it will be calculated as

$$P(A|B) = \frac{P(B|A).P(A)}{P(A)} \quad (7)$$

The probability of 'A' occurring when 'B' has already occurred is calculated as the probability of 'B' occurrence when 'A' already occurred with the actual probability of 'A' occurrence. The probability of actual 'A' occurrences is contained in the probability of 'A' when the probability of 'B' already occurred.

The classifications and predictions can be made using a naïve Bayes classifier, which produces either equal or independent outcomes. The attribute value does not match any other attribute value considered *independent*. The attribute values exactly match each other and are considered to be *equal.*

The Bayes theorem for conditional probability can be measured with the class variable 'y', where 'X' is the parameter.

$$P(y|X)\frac{P(X|y)P(y)}{P(X)} \quad (8)$$

The value of parameter 'X' can be measured with the series of features in 'X' as

$$X= (x_1, x_2, x_3, x_4, \dots \dots . . x_n) \quad (9)$$

The features of the 'X' can be applied for 'x' in a sequence they occur to obtain the values for each attribute in the dataset.

$$P(y|x_1,\dots,x_n)= \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)} \quad (10)$$

The predictors can be used to obtain the exact class labels by applying the values for respective attributes from the dataset to obtain optimal results. The probabilities are identified using the frequency tables along with the similarity measures at every level.

The Bayes theorem for conditional probability can be measured with the class variable 'n', where 'X' is the parameter.

$$P(n|X)\frac{P(X|n)P(n)}{P(X)} \quad (11)$$

The value of parameter 'X' can be measured with the series of features in 'X' as

$$X= (X_1, X_2, X_3, X_4, \dots \dots . . X_n) \quad (12)$$

The features of the 'X' can be applied for 'x' in a sequence they occur to obtain the values for each attribute in the dataset.

$$P(n|x_1,\dots,x_n)= \frac{P(x_1|n)P(x_2|n)\dots P(x_n|n)P(n)}{P(x_1)P(x_2)\dots P(x_n)} \quad (13)$$

The predictors can be used to obtain the exact class labels by applying the values for respective attributes from the dataset to obtain optimal results. The probabilities are identified using the frequency tables along with the similarity measures at every level.

The node stores the occurrence of the events whereas the folds contain the probability that how long the particular event has occurred in the node. The class with the highest probability is the exact prediction from the given dataset. The frequency and the similarity table cover the attribute measures for finding folds in each node.

IV EVALUATION

The most widely used evaluation method in NLP modeling is the Area Under Curve (AUC) method.

It is a classification method to evaluate and solve prediction problems.

A. Area Under Curve Method

Table 2. The Classifiers with actual and predicted values

The area under the curve is a method used for the evaluation of binary classification problems. It has a classifier that classifies the data in terms of

- True Positive Rate (TPR)
- True Negative Rate (TNR)
- False Positive Rate (FPR)
- False Negative Rate (FNR)

1) *True Positive Rate (TPR)*

True Positive Rate (TPR) refers to the number of positive data points that are measured as positive, concerning all positive data points. It is measured as

$$TruePositiveRate = \frac{TruePositive}{False\ Negative + True\ Positive}$$

2) True Negative Rate (TNR)

True Negative Rate (TNR) refers to the number of negative data points that are measured as negative, concerning all negative data points. It is measured as

$$TrueNegativeRate = \frac{TrueNegative}{True\ Negative + FalsePositive}$$

3) False Positive Rate (FPR)

False Positive Rate (FPR) refers to the

$$FalsePositiveRate = \frac{FalsePositive}{True\ Negative + False\ Positive}$$

4) False Negative Rate (FNR)

False Negative Rate (FNR) refers to the number of negative data points that are considered as negative, concerning all negative data points**.** It is measured as

label and the actual values occur in the ‘y’ label respectively.

$$FalseNegativeRate = \frac{FalseNegative}{False\ Negative + TruePositive}$$

B.Precision

Precision is the number of true positive data points that are wrongly considered as positives, concerning all positive data points. It is measured as

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

C.Recall

A recall is the number of true positive data points that are considered as negative, concerning all positive data points. It is measured as

Area Under Curve (AUC) is the binary classification method that ranges from ‘0’ to ‘1’. The sgreatest values between the ‘0’ and ‘1’ will be considered to be the best highest occurrence value ‘h’ among the all-data points.

$$\boldsymbol{Recall = \frac{TruePositive}{TruePositive + FalseNegative}}$$

D. ‘F’ Score

The ‘f’ score is the measure that balances both Precision and Recall methods as a single measure.

$$f = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

The ‘f’ score predicts the occurrences of both the precision and recall values and produces accurate results between them.

V RESULTS AND DISCUSSIONS

The implementations of the NLP Model on the training sets have been made using python and obtained accurate results.

The evaluation table contains the predicted and actual values of TPR, TNR, FPR, and FNR. The predicted values occur in the ‘x’

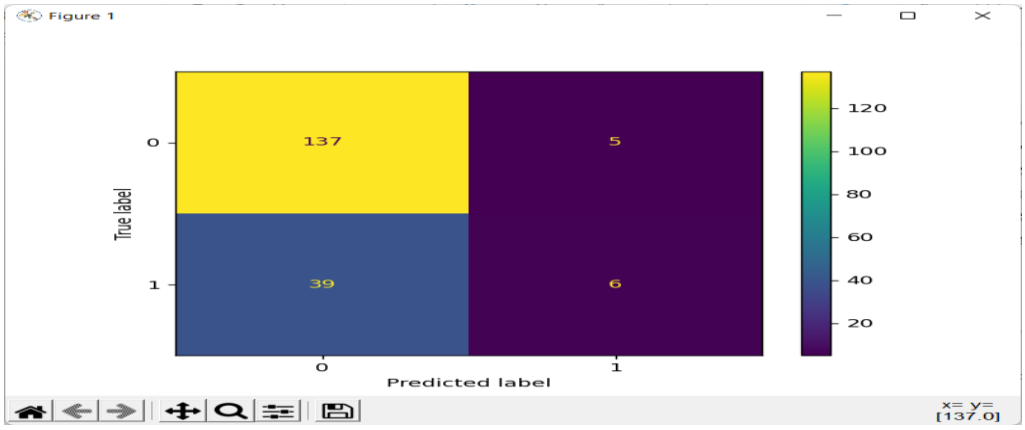

Figure 2. Evaluation table for TPR, TNR, FPR, FNR

The folds of True Positive Rate (TPR) and False Positive Rate (FPR) lie between 0.0 and 1.0. The above graph depicts the curve which connects the occurrence of TPR in the ‘x’, and the FPR in the ‘y’. The prediction of the highest occurrence of TRP and FPR are 0.2 and 0.9 respectively shown in figure3.

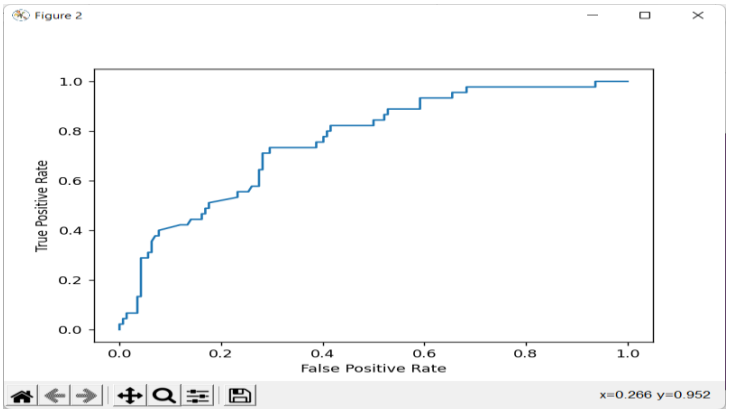

Figure 3. Comparison of TPR and FPR

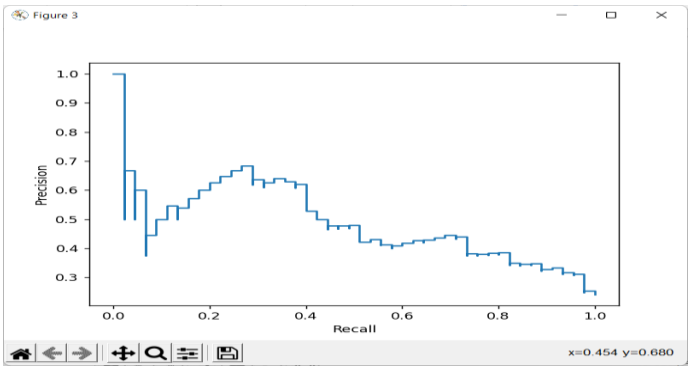

Figure 4. Precision and Recall

The above Figure4. depicts occurrences of precision and recall values which lie between 0.0 to 1.0. The curve connects the occurrence of Precision in the ‘x’, and the Recall in the ‘y’. The highest prediction of Precision and Recall are 0.5 and 0.7 respectively.

The probability measure P(Yes|X) and P(No|X) is measured, using the Naïve Bayes equation to find the probability of ‘yes’ is the occurrence of the text in the folds of each node and the probability of ‘No’ is that the non-occurrence of text in the folds of each node.

Tabel 3.Probability measure for P(Yes/No)

P(Yes|X)	P(yes|X)=P(LTF|yes)*P(GTF|yes)*P(F|yes)*P(yes) =0.31*0.73*0.69*1 =0.16
P(No|X)	P(No|X)=P(P(LTF|no)*P(GTF|no)*P(F|no)*P(no) =0.69*0.59*0.31*1 =0.126

The class with the highest probability is the dataset's exact prediction. The frequency and the similarity table cover the attribute measures for finding folds in each node. The prediction of text gets the highest occurrence of a 'yes' attribute value of '0.16' than the non-occurrence of a 'no' attribute value of '0.126', therefore the given feature is predicted as 'yes'.

V. CONCLUSION

The predictions are made on the real data sets using the python language. The training sets of geographical data used for this analysis are difficult to analyze as of their nature. The latitude and longitude of the geographical area in which most of the word occurs in the tweet are identified. The most efficient and reliable method called the Natural Language Processing (NLP) process applied to the data and predicts the results in terms of probability measure.

REFERENCES

[1] M. Demirci H. Gozde M.C. Taplamacioglu "Fault Diagnosis of Power Transformers with Machine Learning Methods Using Traditional Methods Data", International Journal on Technical and Physical Problems of Engineering" (IJTPE), pp. 225-230, Vol. 13, Issue 49, December 2021.

[2] Chenmeng Zhang, Can Hu, Shijun Xie, Shuping Cao, "Research on the application of Decision Tree and Random Forest Algorithm in the main transformer fault evaluation", Journal of Physics: Conference Series, pp. 1-7, 2021.

[3] Rick Sauber-Cole and Taghi M. Khoshgoftaar, "The use of generative adversarial networks to alleviate the class imbalance in tabular data: a survey", Journal of Big Data, pp. 1-37, 2022.

[4] Terrie Lynn Thompson, "Data-bodies and data activism: Presencing women in digital heritage research", Big Data & Society, July–December: 1–7, 2020.

[5] Sarah Ames and Stuart Lewis, "Disrupting the library: Digital scholarship and Big Data at the National Library of Scotland", Big Data & Society, July–December: 1–7, 2020.

[6] Wen-Tao Wu1, Yuan-Jie Li, Ao-Zi Feng, Li Li, Tao Huang, An-Ding Xu, and Jun Lyu1, "Data mining in clinical big data: the frequently used databases, steps, and methodological models", Military Medical Research (MMR), 2021.

[7] Xin Qiao and Hong Jiao, "Data Mining Techniques in analyzing Process Data: A Didactic", Frontiers in psychology, Vol. 9, November 2018.

[8] Gunther Schuh, Gunther Reinhart. Jan-Philip Prote, Frederick Sauermann,Julia Horsthofer, Florian Oppolzer,Dino Knoll, "Data Mining Definitions and Applications for the Management of Production Complexity", Elsevier, pp. 874-879, 2019.

[9] Bharati M. Ramageri, Maithili V. Arjunwadkar, "Applications of Blockchain Technology in Various Sectors: A Review", International Journal of Future Generation Communication and Networking, pp. 94 – 99, Vol. 13, No. 2, (2020).

[10] Guoguang Rong, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, Mohamad Sawan "Artificial Intelligence in Healthcare: Review and Prediction Case Studies", Elsevier, pp.291-301, 2020.

[11] Yiran Chen, Yuan Xie, Linghao Song, Fan Chen, Tianqi Tang, "A Survey of Accelerator Architectures for Deep Neural Networks", Elsevier, pp. 264-274, 2020.

[12] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, Zhichao Jiang, "Causal Inference", Elsevier,pp.253-263, 2020.

[13] Viktor Schlegel, Goran Nenadic and Riza Batista-Navarro" A survey of methods for revealing and overcoming weaknesses of data-driven Natural Language Understanding", Cambridge University Press, pp. 1-31,2022.

[14] Adam Kovacs, Kinga Gemes, Andras Kornai and Gabor Recski, "Explainable lexical entailment with semantic graphs", Cambridge University Press, pp. 1-24,2022.

[15] Charles Chen, Razvan Bunescu∗ and Cindy Marling, "A semantic parsing pipeline for context-dependent question answering over temporally structured data", Cambridge University Press, pp. 1-25,2022