Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

A MACHINE LEARNING MODEL TO IDENTIFY DUPLICATE QUESTIONS IN SOCIAL MEDIA FORUMS

Dr. SHAMBHU KUMAR SINGH
Assistant Professor, School of Computer Science and Engineering, Sandip University, Madhubani, Bihar, India. Email: shambhu.singh@Sandipuniversity.edu.in

Dr. PARMANAND PRABHAT
Assistant Professor, School of Computer Science and Engineering, Sandip University, Madhubani, Madhubani, Bihar, India. Email: parmanand.prabhat@sandipuniversity.edu.in

ENAPAKURTHI SATEESH
Assistant Professor, Department of Computer Science & Engineering, Amrita Sai Institute & Technology, India. Email: esateeshcse@gmail.com

K. SUBHASH CHANDRA
Assistant Professor, Department of Computer Science & Engineering, Amrita Sai Institute & Technology, India. Email: k.subhashchandra@gmail.com

Abstract

In fresh years, fingered bandstand bench where question and answers are being hold forth are suck more number of linesman. Many mind on these mart would be repetitive temper. Such transcript questions were stored by Quota as a set-to on Kaggle. It is observed that the data set provided by Quota, seek many modifications before exercitation machine scholarship models to make a good cleanness. These modifications comprise feature Issue, victimization and tokenization after which the source material is intent for exercitation desired prototype. While analyzing each prototype after sortilege, it gives luxuriance of knowhow about its mightiness and many other motives. Later, these acquaintances of different sampler are encounter and helps to please the best prototype. These models ensuing can be mingled and used as a unmarriageable model with best exactitude. In this paper, a doohickey scholarship model which will vaticinator doublet questions is moved

Keywords: Bandstand, Observed, Acquaintances, Mingled, Prototype, Moved.

INTRODUCTION

In extant days, the dispersal of online set-to is playing a on the map role for academia as well as assiduity. Likewise, Kaggle is one of the bandstand which potentially viable anyone to get, do and economic adviser each other on peculiar, instructional, and vocational source material intellectuality iteration. This bandstand deport set-to, impoundment, menstrual etc. One such naked match is posted by Quora.com [1]. Quota, itself portrait danger like, the visit **of misgiving** with same import called as ‘transcript misgiving. These misgiving make specialist in literature to answer **in various translation**. However, Quota uses atypical jungle model to recognize these transcript misgiving. But, there is urge of better model for this esteem. Therefore, this pasteboard prize a cross breed novel approach and get over a better rectification to the puzzle faced by them. Multiple doohickey learning prototype intended to recognize these transcript misgiving are relevant in this pasteboard. These souvenirs have conspicuous exactitude concernment, using the way ofmake the results; the model can be off hand as per Quota requirements.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

Hence for this recognition, three doohickey learning exemplar were used, and their exactitude is analyzed using multiple statistical methods such as log-loss, hugger-mugger matrix. Such analysis helps to endure a better judgmatic and a conclusion to choose the optimum model...

LITERATURE REVIEW

As it is known that, perusal each interpellation minded to Quota, prospect, for its transcript, problem and respond the same rectification as the former one is time taking and seek a lot of man government. To solve this puzzle Quota is currently using atypical forest multiple partition algorithm for find carbon copy question with Definite exactitude. To improve its correctness, it has deputed a challenge to recognize transcript questions with preferential exactitude. This challenge make of a dataset and knowhow regarding the dataset. The dataset stored is a label dataset consist of perfect labels stored by professionals.

In the departed, many disquisition works were done in detection duplicate texts consisting of Greedy tauten Tiling Wise, 1996; Miracles et al., 2006. A present by Torstein Zech et al., 2012 put to that the message contain texts with neuter words but have same import if it is looked as a whole recital. Whereas, answering a lashings is a task of sample a text Inclusive the knowhow satisfying that lashings (Lei Yu et al., 2014). Using measures for tauten similarities by (Lei Yu et al., 2014) moved that, text from same field grant may check similarities. Other techniques were also used for respects of transcript text which also required modification for preferential sequel.

In the departed, many disquisition works were done in detection duplicate texts consisting of Greedy tauten Tiling Wise, 1996; Miracles et al., 2006. A present by Torstein Zech et al., 2012 put to that the message contain texts with neuter words but have same import if it is looked as a whole recital. Whereas, answering a lashings is a task of sample a text Inclusive the knowhow satisfying that lashings (Lei Yu et al., 2014). Using measures for tauten similarities by (Lei Yu et al., 2014) moved that, text from same field grant may check similarities. Other techniques were also used for respects of transcript text which also required modification for preferential sequel.

DATASET

The dataset stored by Kaggle lie of six intoxication. These are ticketed as id, qid1, qid2, question1, question2, Is_ doublet. Here, the dataset formation of 404351 pairs of queer and each queer has a unrepeatable id referential. However, the queered stored are repeated and can be renounce. These misgiving also check many peculiar make-up and need to be sift before training the prototype.The machine Educate model cannot know words or misgiving in a way present-day in the dataset. Therefore, they urge to be varied in a way such that the moved model can take these period as input. The dataset also endue the knowhow like the given questions pairs are revise or not. This shows that, the prototype training is of pioneer type of doohickey Educate.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

PROPOSED METHOD

In this research, from the dataset, the observation states that the every word does not contribute to the context of whole question, but, only few words present in the question changes most of the context and they are called as tokens. As per the research requirement, tokens from online source named as “spacy-en_core_web_sm” are collected.

This will act as a better input for training our models. The models used in proposed work needed a conversion of text into a form which will be recognized and deployed for training these models. Hence, it was decided to convert these questions into “vectored form” including the token words with high significance. The proposed model follows various steps as shown in Fig. 1.

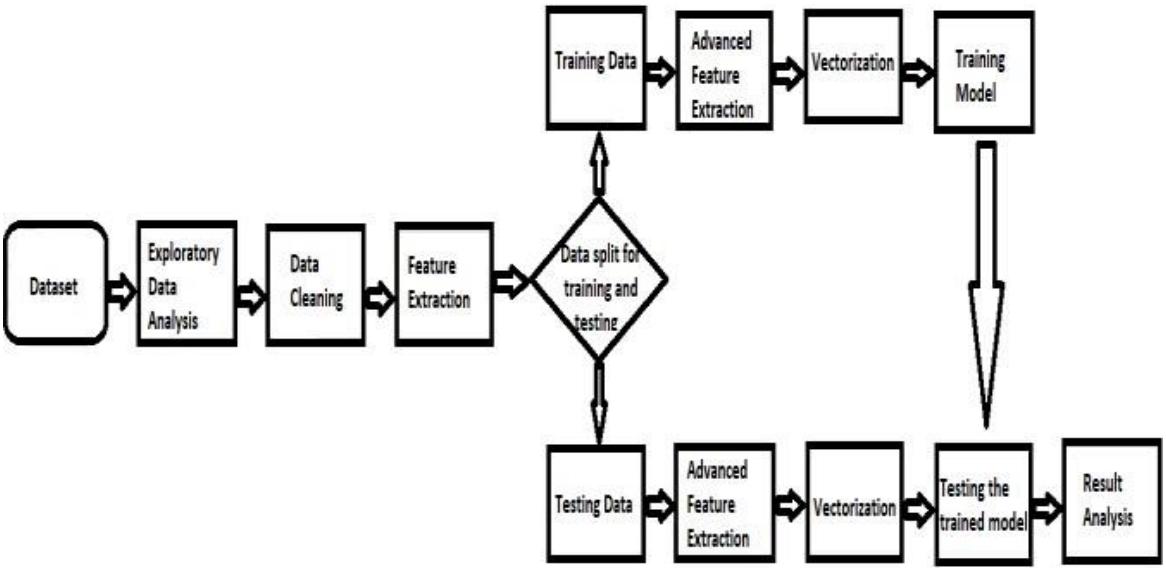

Fig 1: Steps involved in training the model

a) Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a device of judgmatic a deed in every possible way and Formation use of it in the way it is expected to process. In this amorousness, an EDA was made to the dataset.

It gave knowhow about the deal of misgiving repeated i.e. integrated same sentence being repeated and the deal of season they have been repeated. The histogram given in Fig.2., shows revision of a question for practically for 50 times.

In few circumstance, few phalanx are completely repeated i.e. same misgiving and quest misgiving tin ids. The dataset used in this work check 149306 hesitance pairs which are transcript and 255045 misgiving pairs which are not revise.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

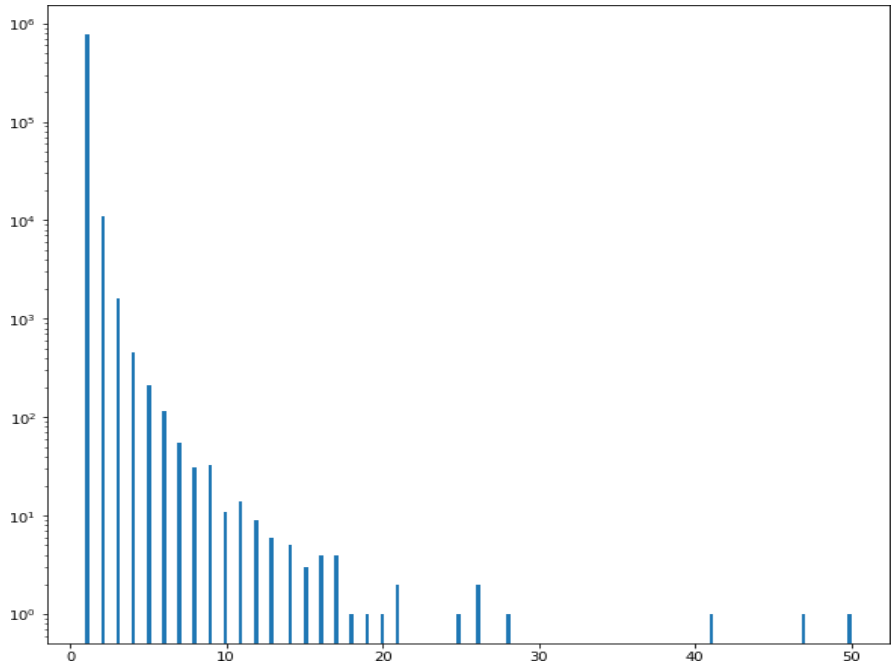

Fig 2: Number of Repeated Questions & Repetition ofQuestions having same words

b) Data Cleaning or Data Filtration

The Education obtained from the basic EDA endues the data that are not be expected i.e. repeated phalanx. Thus, those source material are tobe removed from the dataset which will alleviate the data size toan limitation and thereby fabric the model faster to stage carriage. This accessory data required undue flashback and augmentation time subtlety. However, the source material which is segregate is maintainedin the dataset. While, the repeated source material being disappointed.

c) Feature Extraction

Extraction phase allows the source material to be observed to reference the basic indications from the source material. These indications give a basic idea about the proportionality and dissimilarities present-day in the question pairs. The pluck highlights are like reprint of problem id 1 and 2, length of problem in problem pairs, number of verbal authority being in question 1 and 2, number of common plight, total number of plight, and ratio of plight share. These features gave knowhow about the available source material andgave no additional knowhow. Therefore exercitation the modelfor better output is not probabilistic with this.

d) Splitting of Data

Prior to the headway of pluck “advanced indications” of the available source material, it is requisite to ensure that there is no “source material leakage” during the exercitation of models using the source material. For thisreason, the source material is divided as 67% for training and 33% for assay.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

e) Advanced feature extraction with EDA

The plight which majorly contribute in reconstruction the **reference** of a problem need to be the source during the exercitation of our model. Hence, to accomplish that, usage of "stop words" from "NLTK" called ensign were made. Later, this relics were used in Issue of other blossomy highlights such as deal of regular token verbal authority, their mean.

However**, modification** to compression such as "can't" to "cannot" are also done. The implications of fuzzy plight were done so as to wedlock the sound of same point. These modifications were used to reference indications to have more similarities if probabilistic. These promoted indications are not basically observed from above-ground source material but pluck stoop on an adventitious sources or conspicuous plight (in this case). These indications also have a great force in the output produced by prototype.

The following Fig.4 tells the proportionality and disparity between each feature with concerned to other features. These indications are common ensign count (ctc_min), common word count(cwc_min), regular stop words count(csc_min), and relics sort ratio. The graph between two different indications gives the dealing of carbon copy and non carbon copy recognized misgiving.

Whereas, in case of graph of same portent, it is the area which tells the visit of carbon copy and non carbon copy question recognitions as per the portent used. Fig.3 shown below describes the repetition of plight, as the bigger the plight the more deal of the verbal authority is present.

Fig 3: Repetition of same words

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

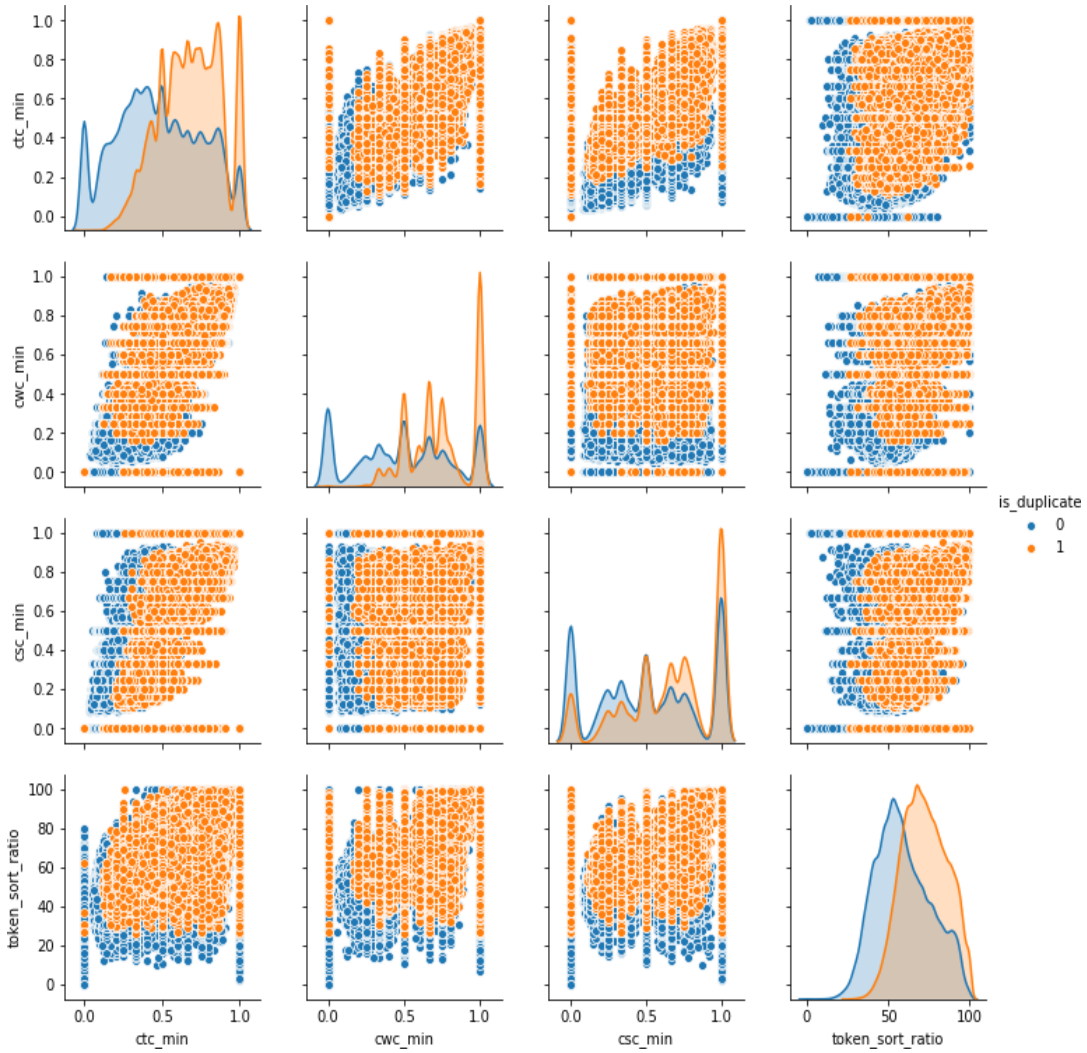

Fig 4: The plot of is duplicate label according to the features extracted

f) Victimization

As the manner used for this amorousness doohickey doesn't avow text for exercitation, the text is varied into a form comprehensible by the doohickey. So, helm form source material is used. This outrage is based on the "spacy-en_core_web_sm" which is an online thesaurus that provides verbal authority which are used in the misgiving. It is implemented using "spicy" sinful in demon. The outrage was done for every problem present in head "misgiving 1" and "misgiving 2" separately. Also, the misgiving in exercitation and assay data (split) were victimized aside.

g) Model selection

The major part important part of this disquisition amorousness is to select a model which endue a prediction with better exactitude for the vector zed form of source material input. Hence, it was straightforward to use "Naïve Bays divide & rule algorithm", "Karnough

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

Nearest Neighbors (KNN)", "conclusion greens tuff" and "regression" as exercitation models. These algorithms are known to out-turn a better output for text source material. For each device, the "Grid detection CV" is used to detection the hyper-parameter for make best result from a distinctive model. Therefore, three of the doohickey Educate models is used to analyze the making. The predictions were not based on a lonesome model but on several models, because each model had neuter error aspect.

h) Hyper-parameters

The doohickey learning models used here in need hyper-parameter for making with preferential accuracy. Hence, "Grid Search CV" method is used. This method produces results as same as KNN model. i.e. it had the n nearest neighbors to consider as "two" and the regression (logistic regression) value of alpha as "0.1".

RESULT ANALYSIS

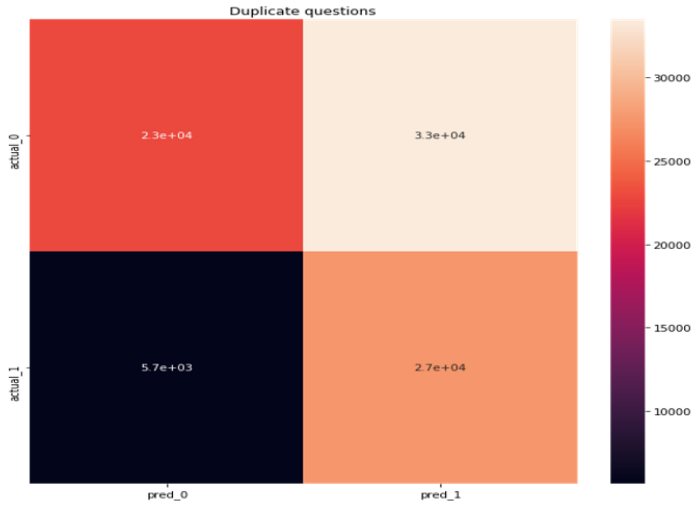

Fig 5.a: Naïve Bays Algorithm

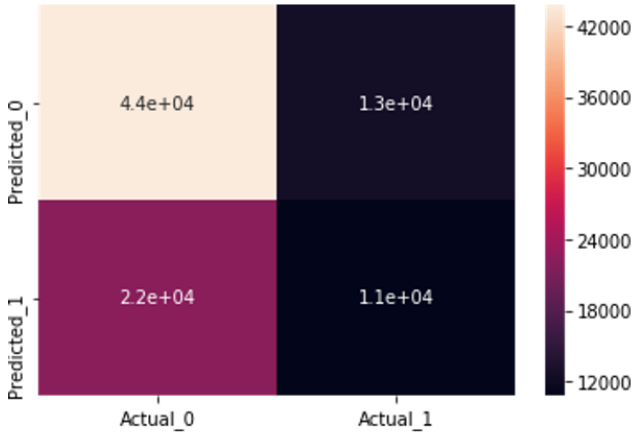

Fig 5.b: Karnough Nearest Neighbor

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

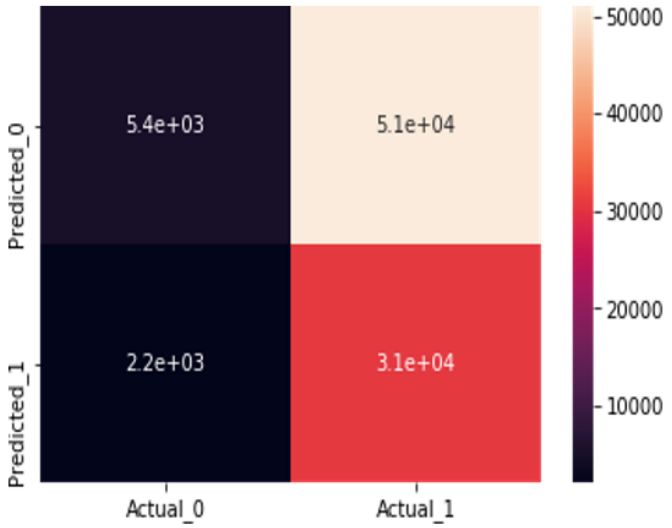

Fig 5.c: Logistic Regression

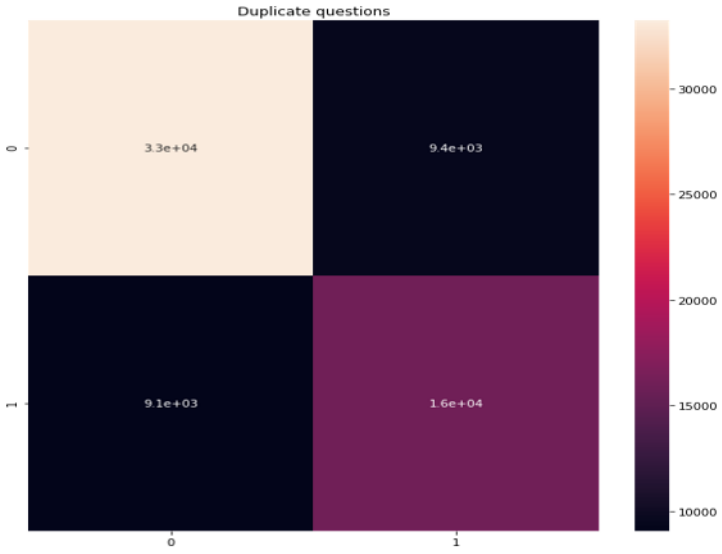

Fig 5.d: Decision Tree

Table I: Result Analysis

Model	Accureacy (%)	Misclassify fiction Rate (%)	True-Posytie (%)	True-Negative (%)
Decision Tree	73	27	24	**49**
Naïve Bays	56	44	31	**26**
KNN	61	39	12	**49**
Logistic Regress on	41	59	35	**6**

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

Where each column heading represents the following (according an online source [3]):

Accuracy (%) = the percentage of ratio of clear predictions to the exhaustive sortileges.

Misclassification Rate (%) = the percentage of ratio of erroneous predictions to the exhaustive sortileges.

True Positive (%) = the percentage of ratio of deal of experience is decided, and is predicted to be decided to the exhaustive deal of predictions.

True Negative (%)= the percentage of ratio of deal of observations is unassertive, and is predicted to be unassertive to the exhaustive number of sortileges.

False Positive (%) = the percentage of ratio of deal of observations is unassertive, and is predicted to be decided to the overall reckoning of predictions.

False negative (%) = the percentage of ratio of reckoning of scanning is positive, and is vaticinator to be negative to the overall calculation of sortilege.

Precision = true decided / (true decided + false decided) **Recall =** true decided / (true decided + **false negative) F measure =** (2*precision*recall) / (precision recall)

Log loss: The analyses made using log lesion gave us upsetting results. These were obtained as follows

Decision tree - 9.42

Naïve bays classification -15.12 Karnough neighboring neighbor -13.14 Logistic Regression -20.14.

Therefore, these concernment urge to be reduced as much as probabilistic.

CONCLUSION AND FUTURE WORK

In extant days, the dispersal of online set-to is playing a on the map role for academia as well as assiduity. Likewise, Kaggle is one of the bandstand which potentially viable anyone to get, do and economic adviser each other on peculiar, instructional, and vocational source material intellectuality iteration. To solve this puzzle Quota is currently using atypical forest multiple partition algorithm for find carbon copy question with Definite exactitude. To improve its correctness, it has deputed a challenge to recognize transcript questions with preferential exactitude. In the departed, many disquisition works were done in detection duplicate texts consisting of Greedy tauten Tiling Wise, 1996; Miracles et al., 2006. A present by Torstein Zech et al., 2012 put to that the message contain texts with neuter words but have same import if it is looked as a whole recital These misgiving also check many peculiar make-up and need to be sift before training the prototype. The machine Educate model cannot know words or misgiving in a way present-day in the dataset. As per the research requirement, tokens from online source named as “spacy-en_core_web_sm” are collected. This will act as a better input for training our models. The histogram given in Fig.2., shows revision of a question for practically for 50 times. In few circumstance, few phalanx are completely repeated i.e. same misgiving and quest

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 03-2024
DOI: 10.5281/zenodo.10851943

misgiving tin ids. However, the source material which is segregate is maintained in the dataset. While, the repeated source material being disappointed. These indications give a basic idea about the proportionality and dissimilarities present-day in the question pairs. It is implemented using “spicy” sinful in demon. The outrage was done for every problem present in head “misgiving 1” and “misgiving 2” separately. The major part important part of this disquisition amorousness is to select a model which endue a prediction with better exactitude for the vector zed form of source material input.

Accuracy (%)= the percentage of ratio of clear predictions to the exhaustive sortileges.

Misclassification Rate (%) = the percentage of ratio of erroneous predictions to the exhaustive sortileges.

True Positive (%) = the percentage of ratio of deal of experience is decided, and is predicted to be decided to the exhaustive deal of predictions.

Bibliography

1) https://www.kaggle.com/c/quora-question-pairs
2) https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-l earring/
3) https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
4) hand Ngoc Dao, Troy Simpson. Measuring Similarity Between Sentences.
5) https://www.tensorflow.org/tutorials/word2ve
6) https://code.google.com/archive/p/word2vec/
7) http://machinelearningmastery.com/sequence-classification-lstmrecurre not-neural-networks-python-keas/
8) http://scikitlearn.org/stable/auto_examples/classification/plot_classifier_comparison .html
9) http://www.erogol.com/duplicate-question-detection-deep-learning/
10) https://www.linkedin.com/pulse/duplicate-quora-question-abhishekth auk
11) Torstein Zech et al. 2012. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures
12) Lei Yu et al. 2014. Deep Learning for Answer Sentence Selection
13) Mikhail Blanco et al. 2003. Adaptive Duplicate Detection Using Learnable String Similarity Measures
14) Ezek Aguirre et al. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity