iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35575

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Understanding Question Pair Similarity

Shekkari Akhil [1], Kamarju Sai Anurag[2], Gogi Vishwanath[3], Nyatha Vinay Kumar[4]
[1, 2, 3]*Student,* [4]*Assistant Professor, Electronics and Computer Engineering,*
Jawaharlal Nehru Technological University Hyderabad (JNTUH), India

Abstract- One of the most important areas where the Natural Language Process of Machine Learning may help is determining if two questions are similar. The model we create can instantly detect if a question is similar to one that has already been posed. To find the underlying patterns in our data, we'll do a complete Exploratory Data Analysis. Based on our observations, we will do feature engineering. We'll try out a few different modelling strategies to determine which one works the best and keeps the greatest outcomes.

Keywords - Underlying Patterns, Exploratory Data Analysis, Feature Engineering, Modelling Strategies

I. INTRODUCTION

There are basically many steps for getting the desired output and it is totally different from the existing system where the similarity is shown only based on the no of words. Whereas our existing system shows the similarity based on the context meaning or sentence meaning this can be done through complex process starting from collecting the data and after collecting it cleaning the data, followed by analysing the data and feature engineering, hyperparameter tunning, modelling and finally evaluation, result this is totally an outward process but internally there is a lot of work involved including the coding part and usage of different models.

II. LITERATURE SURVEY

In the existing system, ssimilarity between two sentences is determined by number of common words in both the sentences. The existing system does not imply semantic meaning of the sentences. As a result, we might not get the complete picture of both the sentences. This type of count in similar words in two sentences is known as BLUE Score.

In **Support vector Machine classifier**, we will build a hyper plane in the support vector machine algorithm to classify data points in such a way that the distance between both support vectors is as small as possible. In other words, we're looking for a hyperplane that maximises the margin. Hyper planes that are parallel to our original hyper plane and hit extreme locations are called support vectors.

In **Logistic Regression**, we assume that the data is almost linearly separable or totally linearly separable in logistic regression. In logistic regression, we select a hyperplane with the shortest distance between it and the datapoints.

III. IMPLEMENTING QUESTION PAIR SIMILARITY

We will use machine learning model which is Decision Trees to determine the similarity of the sentences. In other words, we take semantic meaning into account. Here we are implementing our machine learning models with advanced algorithms so as to get better accuracy. Here we used GBDT model.

- Data Collection
- Data Cleaning
- Data Analysis
- Feature Engineering
- Performance Metrics
- Modelling

The data set is taken from Kaggle. It has 404,290 rows and 6 columns. The columns are id, qid1, qid2, question1, question2, is_duplicate.id, qid1, qid2 are numerical features. question1 and question2 are text features. is duplicate target variable here we remove all the punctuation and all sorts of special characters and numbers. We make the text plain so as to pass it to model

There are different ways to handle null values. We cannot leave null values in data set as it creates problems later on. Depending on No. of null data points we can either drop it or impute new values using mean, median and mode.

The data set consists of 404,290 rows in it. Now we have to make train and test data in the ratio of 70 – 30 percent. Then we again split training data for creating cross validation dataset. When splitting the data, we have to be cautious. There are nine cancer classes in the data set. So even distribution of nine classes should be present in training data, testing data and cross validation data. If the distribution of cancer class is not even then models, we train are bound to fail in classifying the data.

©IJRASET: All Rights are Reserved

Feature Engineering is one of the most important parts of machine learning. Feature Engineering lets you create new features which would help in increasing accuracy of the model. Here we have created many different features like word share, length of words, word total, word common etc., some analysis of created features are:

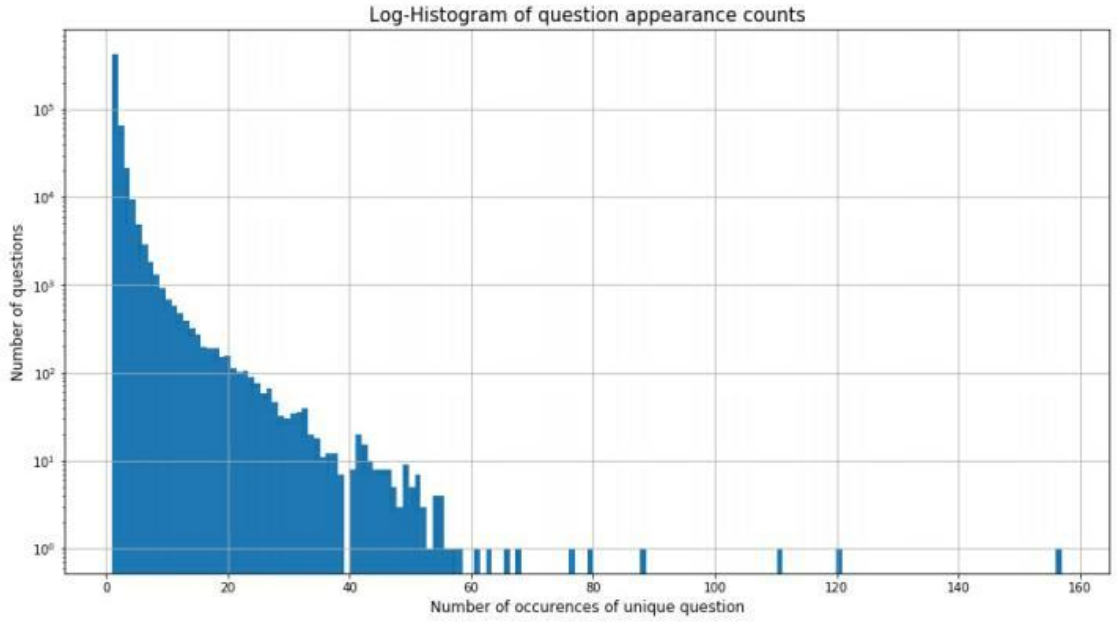


Fig1 represents the log-histogram of question appearance counts

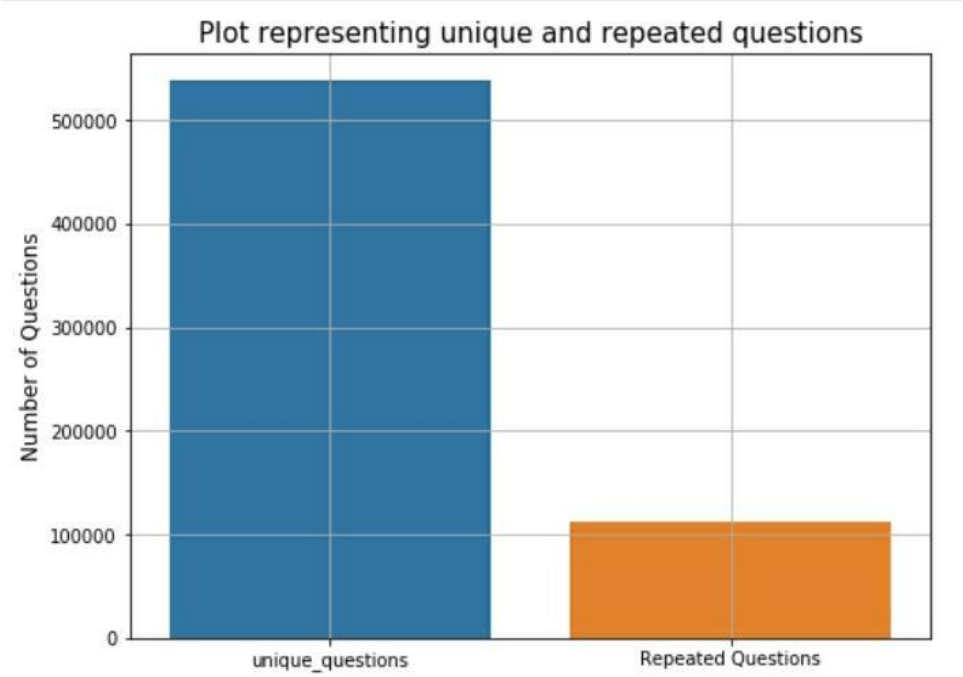


Fig2 describes the no of unique questions and repeated questions

©IJRASET: All Rights are Reserved

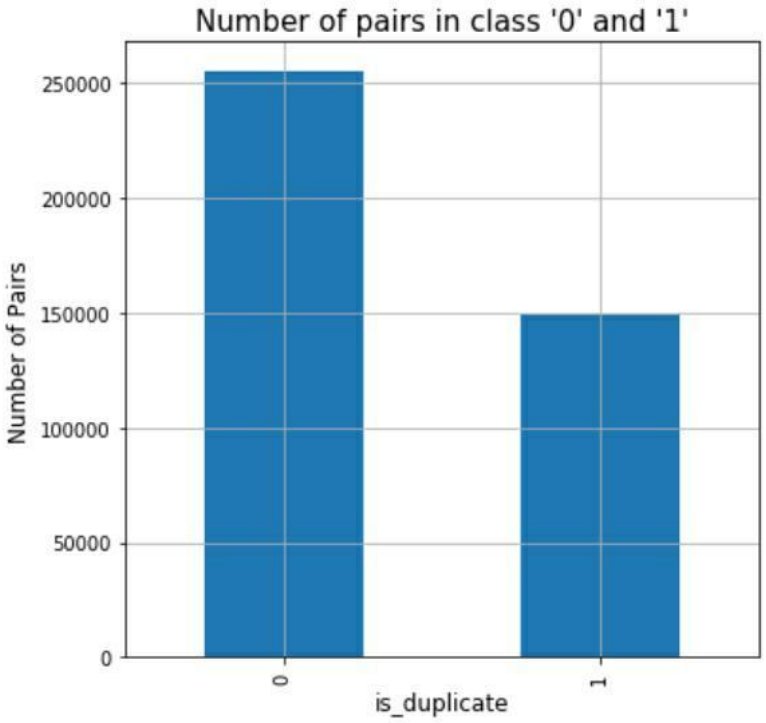


Fig3 Pairs of duplicate questions

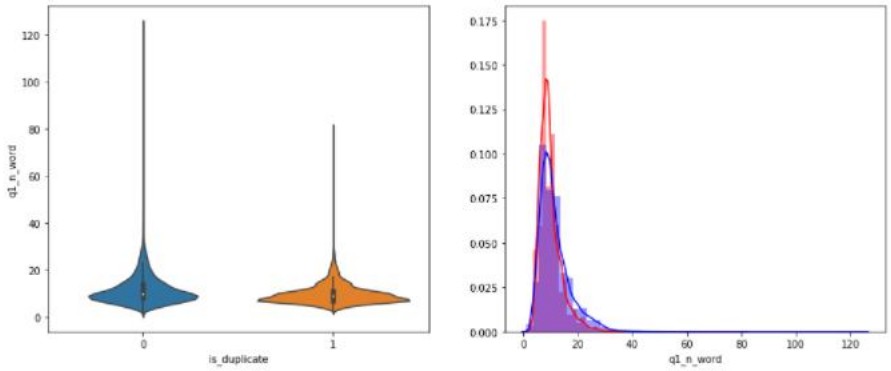


Fig4 Violin plots

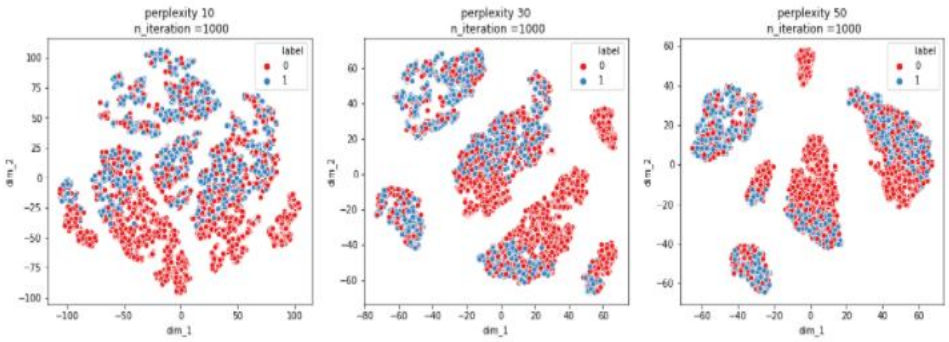


Fig5 T-SNE visualization

©IJRASET: All Rights are Reserved

A. Performance Metrics

The model's efficiency is measured using performance metrics. In machine learning, there are a variety of performance indicators, each of which has its own approach to model evaluation. The most crucial phase in the machine learning process is selecting an appropriate performance metric. The type of dataset and whether it is balanced or imbalanced influence the performance metric chosen.

B. Log loss

It's also referred to as logistic loss. It calculates the probability values of each data point in each class. The log loss has values ranging from zero to infinity; the lower the log loss, the better the model. Log loss is the most appropriate performance metric for interpreting the results of machine learning models where probability is important. The log loss formula is as follows:

$$logloss = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij}\log(p_{ij})$$

Fig 6 log loss

where:

N = Number of data points

M = Number of classes

yij = Actual output value (1 for correct class, 0 for incorrect class)

P(yij) = Probability of data point belonging to class

C. Confusion Matrix

For each data point, a confusion matrix is a two-dimensional representation of the predicted and actual value of outputs. When the data set is unbalanced, this method is utilised (unequal distribution of output variables). Actual values are represented by the rows in the confusion matrix.

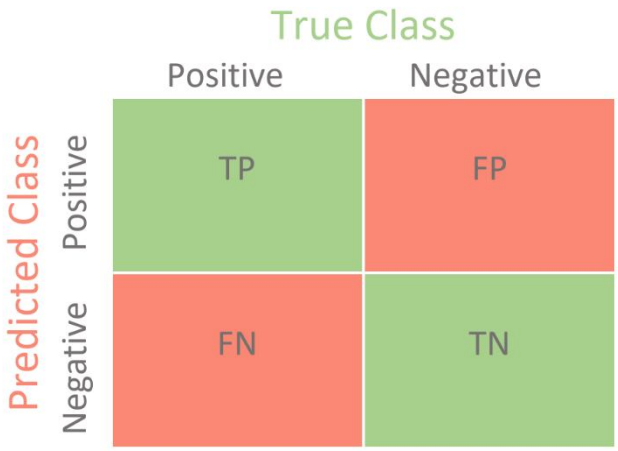


Fig 7 Confusion Matrix

We need to create a model in which the confusion matrix's diagonal values are significantly higher than the non-diagonal ones. Because the expected and actual values are the same, diagonal elements are used. The term "non-diagonal values" refers to the

©IJRASET: All Rights are Reserved

difference between predicted and actual values. As a result, the lower the value of non-diagonal elements, the better our model will be.

D. Precision

Precision is the number of correctly predicted values over the number of all predicted values.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Fig 8 Precision

E. Recall

Recall is the number of correct values over the number of values that should have been returned.

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

Fig 9 Recall

F. Model (XGBoost)

The distance between the query data point and its GBDT k nearest number of neighbours will be calculated using the GBDT method. In the GBDT algorithm, K is the hyper parameter. The value of K can be anywhere between one and infinity. It is preferable to use an odd number for K.

The Euclidian distance, Manhattan distance, or Minkowski distance is used to calculate the distance between the query point and its neighbours. We utilise Euclidian distance in most circumstances, but if we utilise L1 regularisation in logistic regression, we can utilise Manhattan distance. The generalised version of Manhattan and Euclidian distance is the Minkowski distance.

Because the confusion matrix's diagonal members have high values, we can conclude that the classifier is successful in classification. The diagonal elements' darker colours of green reflect the nine classes' higher precision values. Class nine points have been accurately classified 100 percent of the time. With 0.47 percent, class five is the least popular. All classes are recalled by the darker black colours in diagonal elements. They have a high recall, indicating that the data.

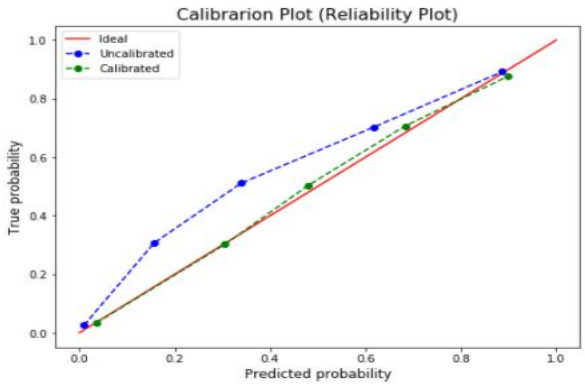


Fig10 Calibration Plot

```
4 pairs are left to be check to tune the hyperparamter:
Training and calibrating for: n_estimators = 400, max_depth = 8
Training and calibrating for: n_estimators = 400, max_depth = 10
Training and calibrating for: n_estimators = 500, max_depth = 8
Training and calibrating for: n_estimators = 500, max_depth = 10
```

Fig 11 Training and Calibration

©IJRASET: All Rights are Reserved

IV. RESULTS & OUTPUTS

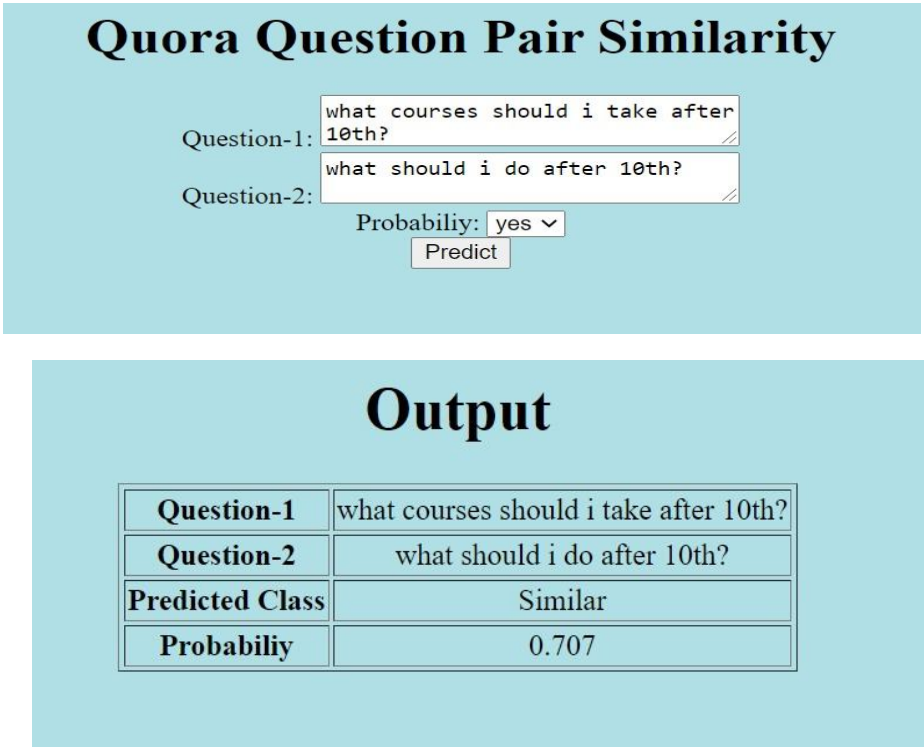


Fig 12 From the figures we can observe that these two questions are similar with the percentage of 70

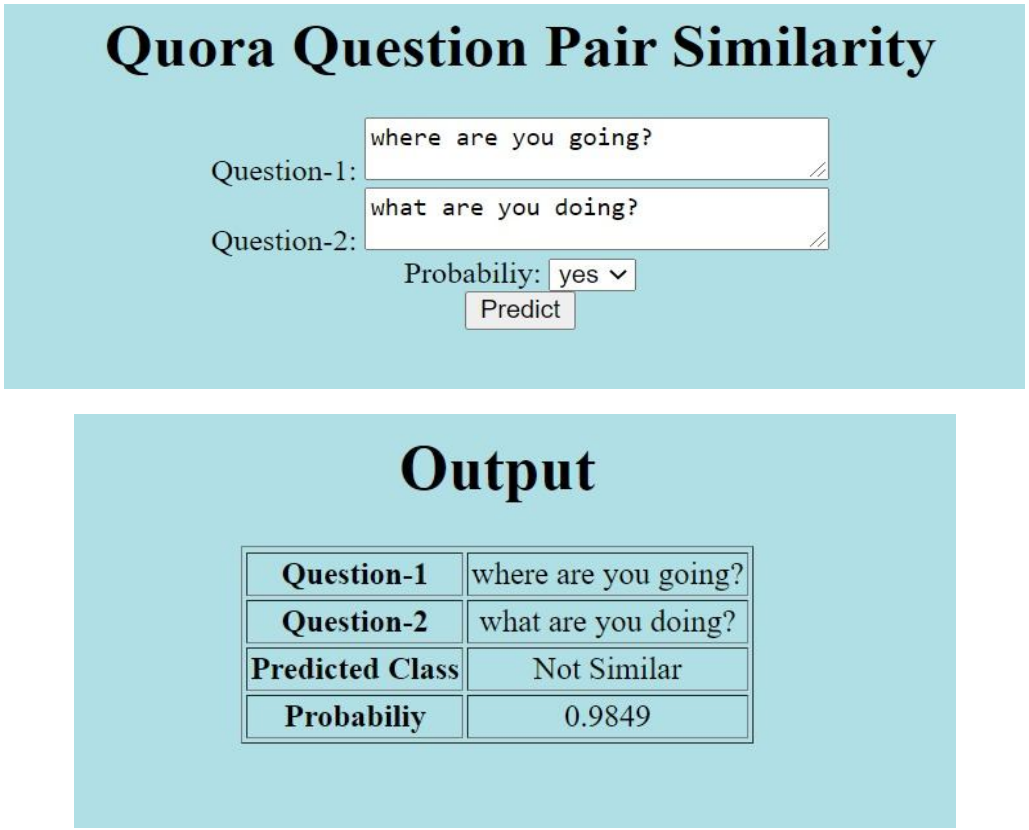


Fig13 From the figures we can observe that these two questions are not similar with the percentage of 98

©IJRASET: All Rights are Reserved

V. FUTURE WORK

This model can be implemented with much richer algorithms which may take birth in future. This can be implemented in IOT for better understanding between individuals and machines given the wide range of uses of this in future.This can be implemented in acro sector, health and also increasing security

VI. CONCLUSION

Finally, even though our application is not real-time, new questions come in every second, and Quora may want to retrain their models on a regular basis. This isn't a set-it-and-forget-it strategy. This means that training must be completed in a fair amount of time, and unlike random forests, GBDTs are not very parallelisable. With a Classifier like this, we can help many major platforms like Quora, Stack Overflow etc., which would help answer duplicate questions saving lot of fortune to both sides. These models in future would be a lot of help.

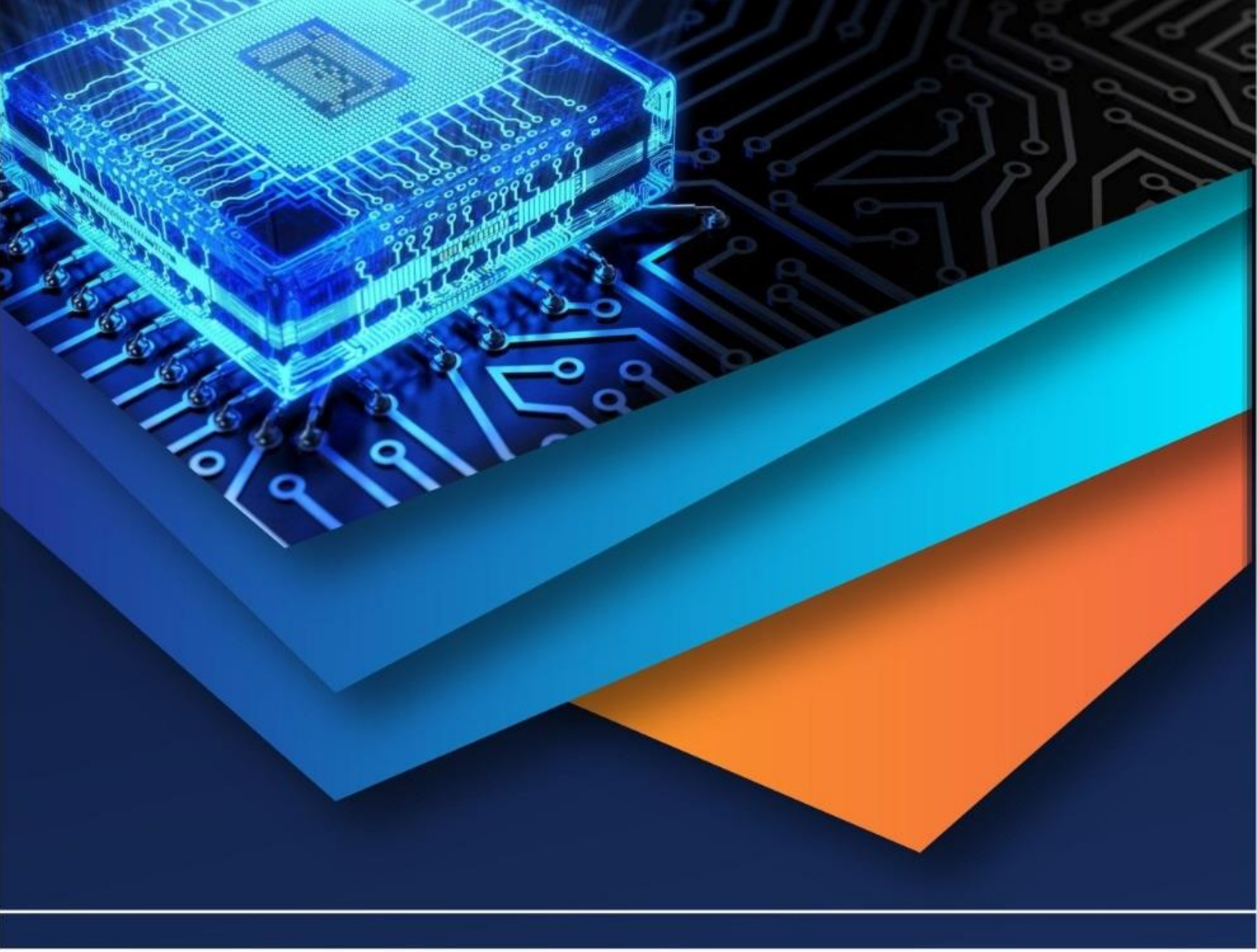
VII. ACKNOWLEDGMENT

We owe a great many thanks to a great many people who have helped and supported us throughout this project, which would not have taken shape without their co-operation. Thanks to all. We express our profound gratitude to Dr.T.Ch.Siva Reddy, Principal and indebtedness to our management Sreenidhi Institute of Science and Technology, Ghatkesar for their constructive criticism.
We would like to specially thank our beloved Head of Department, ECM, Prof. Dr. D.Mohan, for his guidance, inspiration and constant encouragement throughout this research work. We would like to express our deep gratitude Mrs. N.Swapna, Associate professor of ECM and Mr. M. Nanda Kumar, Associate professor of ECM (Project Coordinators) and Mr.N.Vinay Kumar, Assistant professor of ECM (Internal Guide), For their timely guidance, moral support and personal supervision throughout the project.
These few words would never be complete if we were not to mention our thanks to our parents, Department laboratory, staff members and all friends without whose co-operation this project could not have become a reality

REFERENCES

[1] McCreery, C., Katariya, N., Kannan, A., Chablani, M., & Amatriain, X. (2019). Domain-Relevant Embeddings for Medical Question Similarity. *arXiv preprint arXiv:1910.04192*.
[2] Burke, Robin D., Kristian J. Hammond, Vladimir Kulyukin, Steven L. Lytinen, Noriko Tomuro, and Scott Schoenberg. "Question answering from frequently asked question files: Experiences with the faq finder system." *AI magazine* 18, no. 2 (1997): 57-57.
[3] Acitelli, L.K., Douvan, E. and Veroff, J., 1993. Perceptions of conflict in the first year of marriage: How important are similarity and understanding? *Journal of Social and Personal Relationships*, *10*(1), pp.5-19.
[4] Bhattacharyya P, Garg A, Wu SF. Analysis of user keyword similarity in online social networks. Social network analysis and mining. 2011 Jul 1;1(3):143-58.
[5] Deudon, M. (2018). Learning semantic similarity in a continuous space. In *Advances in neural information processing systems* (pp. 986-997).
[6] Achananuparp, P., Hu, X., Zhou, X. and Zhang, X., 2008, April. Utilizing sentence similarity and question type similarity to response to similar questions in knowledge-sharing community. In *Proceedings of QAWeb 2008 Workshop, Beijing, China* (Vol. 214).

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)