ASSOCIATION OF IDENTICAL PAIRS USING NATURAL LANGUAGE PROCESSING

[1]Saravanan Alagarsamy, [2]Kartheeban Kamatchi, [3]Mehta Maharshi, [4]Nilesh Nirav, [5]Moksh Kaushal

ABSTRACT--- *Question duplication is the serious issue experienced by question and answer discussion forum like Quora, Stack-flood, Reddit, and so on. Answers get divided across various adaptations of a similar inquiry because of the repetition of inquiries in these gatherings. In the end, this outcome in absence of a reasonable pursuit, answer weakness, isolation of data and the lack of reaction to the examiners. The copied questions can be identified utilizing Machine Learning and Natural Language Processing. Dataset of in excess of 400,000 inquiries sets gave by Quora are preprocessed through tokenization, lemmatization and evacuation of stop words. This pre-handled dataset is utilized for the element extraction. Fake Neural Network is then planned and the highlights thus removed, are fit into the model. This neural system gives exactness of 86.09%. More or less, this examination predicts the semantic fortuitous event between the inquiry sets removing profoundly prevailing aspects and consequently, decide the likelihood of inquiry being copy.*

Keywords--- *Nature Language Processing**,** Vector Space Modeling, Artificial Intelligence.*

I. INTRODUCTION

More than 100 million individuals visit Quora consistently and in excess of 14 million inquiries have been posted up until now. Along these lines, there is a high possibility that numerous individuals pose comparable inquiries that might be in various structures. This is an extreme issue and consequently, Quora distributed its dataset without precedent for Feb 2017. This dataset comprises of 404,290 inquiry matches alongside the is duplicate parameter. Assortment and representation of the dataset are done before further preparing the dataset. Preprocessing of dataset comprises tokenization, stemming and expulsion of stop words.

All the inquiry sets are then changed over into vectors. Standardized highlights, fluffy fuzzy parameters, TFIDF proportion, word-share proportion, slant variables, and vector separates between the sets of inquiries are

[1] *Assistant Professor, Department of Computer Science and Engineering. Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, Tamilnadu, India,a.saravanan@klu.ac.in*

[2]*Associate Professor, Department of Computer Science and Engineering. Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, Tamilnadu, India, k.kartheeban@klu.ac.in*

[3] *Department of Computer Science and Engineering. Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, Maharshi108@gmail.com*

[4] *Department of Computer Science and Engineering. Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, Niravklu2016@gmail.com*

[5] *Department of Computer Science and Engineering. Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, Mokshkaushal0@gmail.com*

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

determined. Highlight designing includes dealing with 300 measurements; utilizing the Google News vector, which is prepared on approximately 100 billion words from Google News dataset. Highlights are separated from the inquiry matches and afterward, a neural system comprising of five shrouded layers is planned. The most noteworthy precision of 86.09% is accomplished in 16 ages.

II. LITERATURE REVIEW

The Semantic coordinating of sentences beforehand has concentrated on consistent surmising dependent on the Stanford Natural Language Inference Corpus. Alagarsamy et al. [1] concentrated on word-by-word consideration strategies utilizing LSTMs. The Semeval challenge, committed to semantic comparability, was the principal to incorporate an assignment explicitly on question similitude. Copy Question Detection falls under the more extensive assignment of semantic content comparability (STS), which has been the subject of the Semeval challenges since 2012. Early work to identify the similitude between sentences utilized physically built highlights like word cover alongside customary Artificial Intelligence (AI) calculations like Support Vector Machines [2].

Neural Network approaches have been best in class in a wide scope of NLP assignments. Siamese neural system comprising of two sub-systems joined at their yields was proposed. [3]. Despite the fact that the Siamese engineering is lightweight and simple to prepare, there is no impact of connection between the parameters, which may cause data misfortune. Along these lines, to understand the impediments of the Siamese system, the Compare-Aggregate model was proposed [4], which catches the collaboration between two sentences [5]

Information science engineers at Quora as of late discharged an open dataset of copy addresses that are utilized to prepare copy question identification models. Research at the Department of Software engineering, Stanford University and New York University was an extraordinary wellspring of information for us to plunge into. Assessment of a model for separating different highlights was done which incorporates the idea of fluffy fuzzy and vector separations of the writings too. After made a deep survey, there is need of effective mechanism for avoiding the duplication process. The proposed mechanism using the NLP is used as perfect tool to avoid the duplication occurring in the Quora [6].

III. METHODOLOGY

The Natural Language Processing (NLP) is used as a effective tool in the combination of Artificial Intelligence (AI) for semantic analysis, parsing. The proposed method used rule based statically NLP. The step by step execution process of proposed method is discussed below.

Step 1. Data Extraction

Dataset used for the validation is obtained Quora, which contains a complete number of 404,290 legitimate inquiry sets. The dataset is organized as following segment marks: "id", "qid1", "qid2", "question1", "question2" and” is duplicate". Also, the test dataset contains absolutely 2,345,796 inquiry matches yet with no " is duplicate" name. The extracted data is given for the pre-processing stage [7].

Step 2. Data Processing and filtration process

Removing of stop words, tokenization, normalization and stemming are performed at first. Column "id" is dropped since it has no use in the prediction of duplicate question. Similarly, question mark (?) and all the stops words that act as outliers in the dataset are removed. Here is the list of words having high TFIDF score.

Step 3. Data Description

Insights of the datasets are dissected plotting histogram, which gives the proportion of copy question sets. Description of data set is discussed detail in the experimental results section.

Step 4. Relation between two questions

Relationship on word share proportion and TFIDF word share is plotted as appeared beneath which shows that an expansion in the word offer and TFIDF share builds the likelihood of the inquiry being a copy.

Step 5. Feature Extraction

Since Machine Learning models don't work straightforwardly with content, [4] we convert inquiries into numbers or the number of vectors. This is done through Vector Space Modeling (VSM). VSM can to a great extent be actualized by means of the accompanying systems [8]**.**

Step 6. Feature Engineering

Highlight designing includes separating data from the given dataset. Highlights are partitioned into four classes. This includes the working of fundamental NLTK arithmetic, fluffy fuzzy parameters, Word Mover Distance and vector separation [9].

Step 7. Classification of duplicate words

Finally the classification of process is performed. Duplication of word are indentified and words with duplication are classified on the result from NLP.

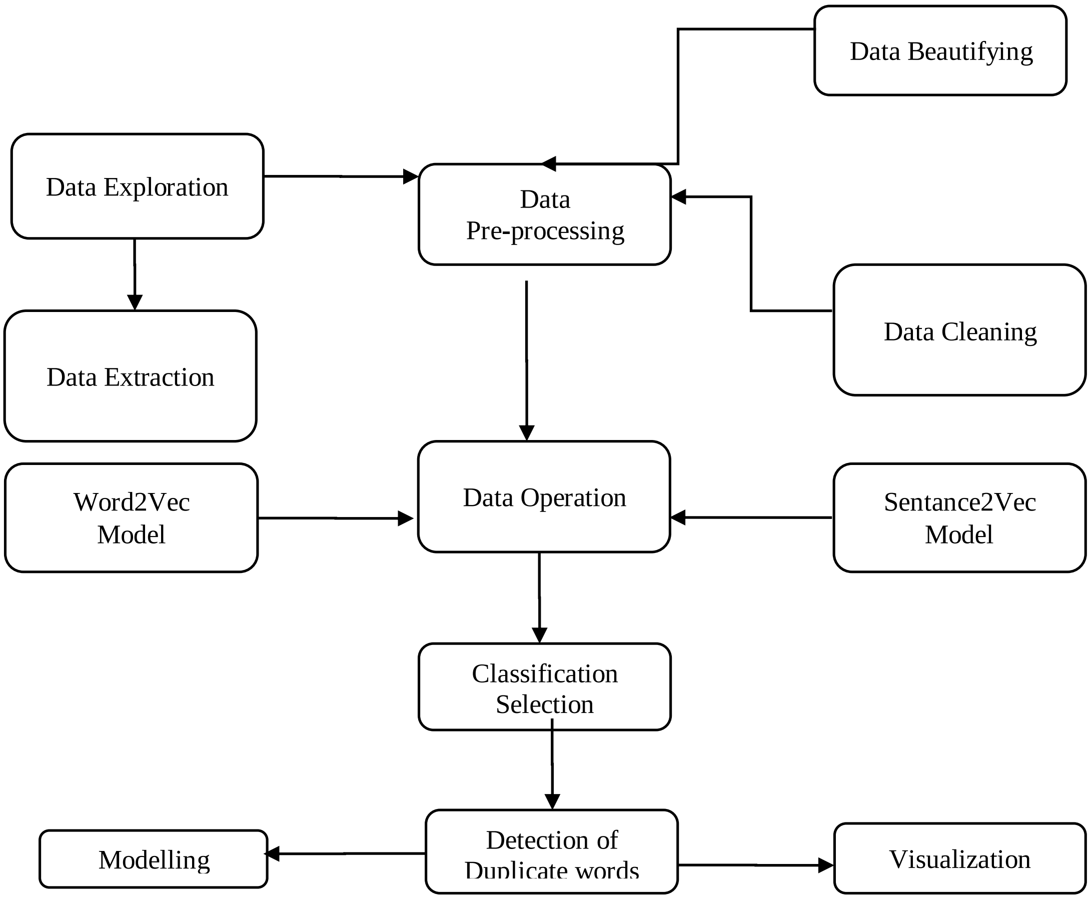

Figure 1: Block diagram of Proposed System

IV. STUDY RESULTS, SUMMARY AND CONTRIBUTION

In this study, data collected from Quora is preprocessed. First, the weights are assigned to each question for identification purpose. Then the length of word is calculated. Finally the detecting the duplicate word in the forum is identified and classification of new and duplication words are identified by using NLP technique.

Table 1: below summarizes questions used for the validation of proposed technique using NLP.

id	qid1	qid2	question1	question2	is_duplicate	freq_qid1	freq_qid2	q1len	q2len
0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0	1	1	66	57

1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0	4	1	51	88
2	5	6	How can I increase the speed of my internet co.	How can Internet speed be increased by hacking...	0	1	1	73	59
3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0	1	1	50	65
4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0	3	1	76	39

The Figure. 2 clearly indicate the analysis of various questions. More than 50000 questions have been considered .Unique and repeated questions are separated effectively by the suggested technique.

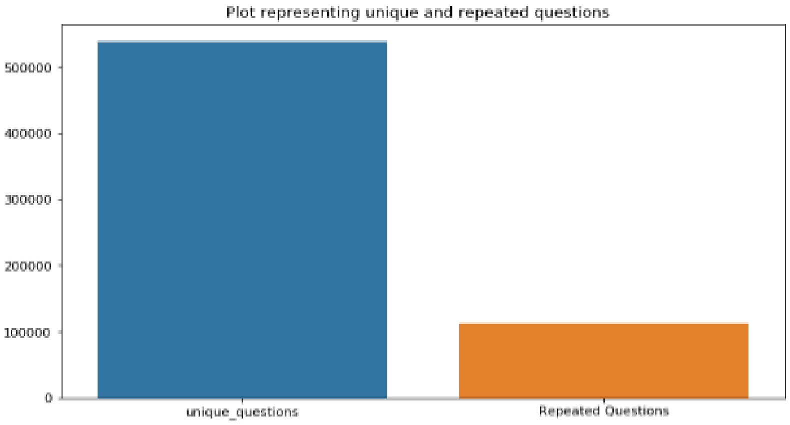

Figure 2: Classification of unique and duplicated questions

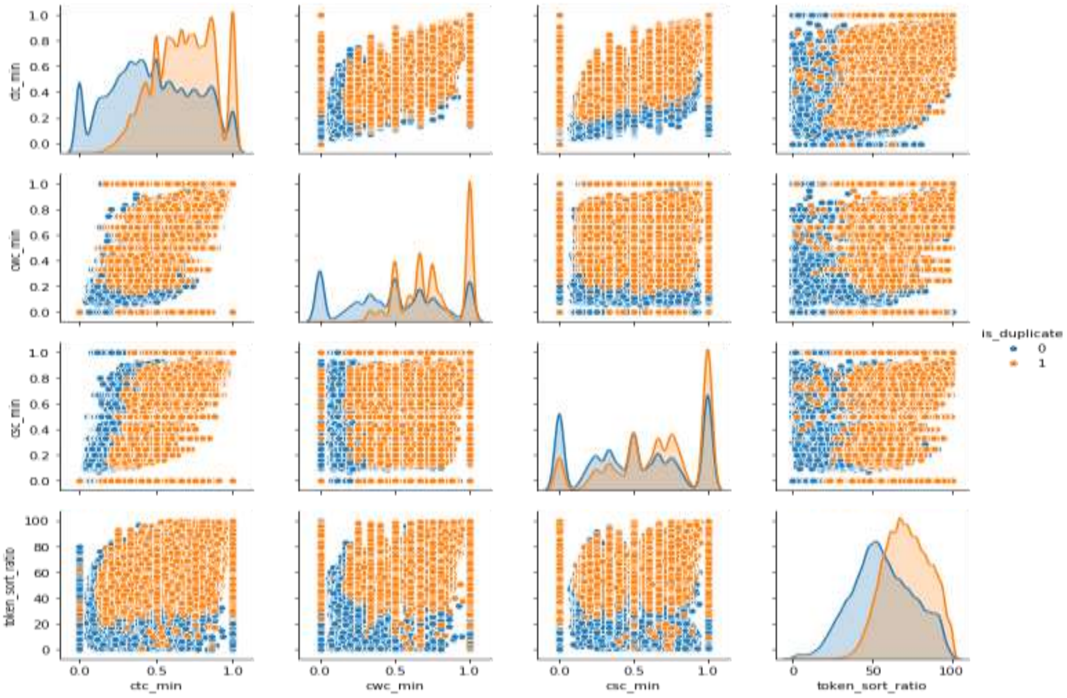

Figure 3: Distribution of weights assigned for new and duplicate questions

The figure. 3 represent the weights assigned for original and duplicate questions. The weight 0 is assigned for the new type of questions and 1 is assigned for the duplication type of questions. Collection used for the evaluation is mixed of both new and duplicate questions. Red dot in the figure indicates the duplicate questions and blue dot represent the new questions [10].

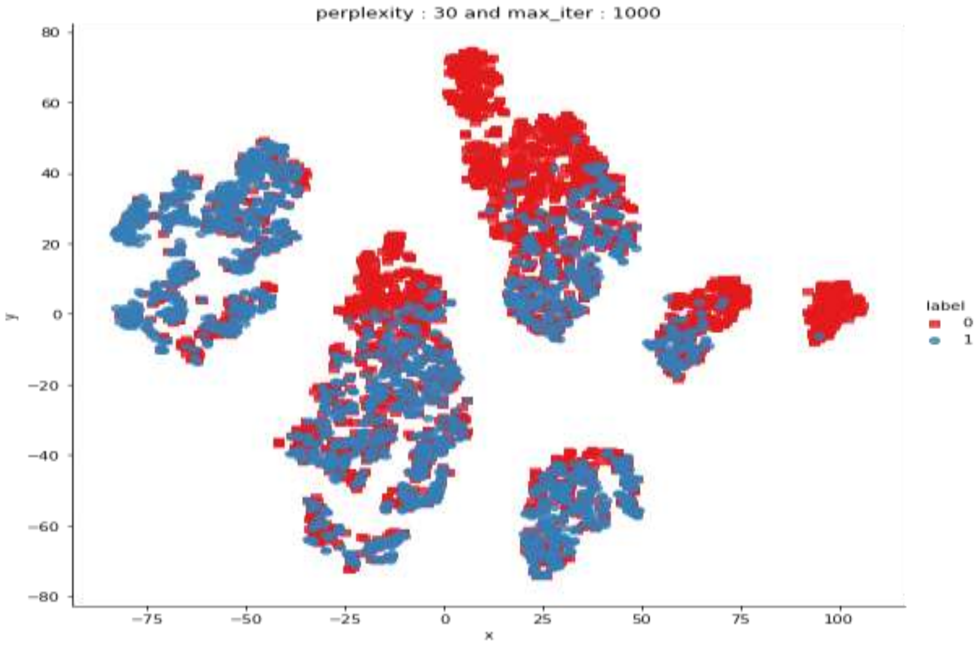

Figure 4: Accuracy calculation for the proposed system.

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

The suggested technique produces 86.09% exactness for identifying the new and duplicate words. Experimental results clearly show the effectiveness for prediction. Figure 5 represents the distribution of token for new and duplicate questions .In the forum, most of the questions are repeated and giving the response for the questions is time of waste .token ratio for distribution is shown in the figure.5

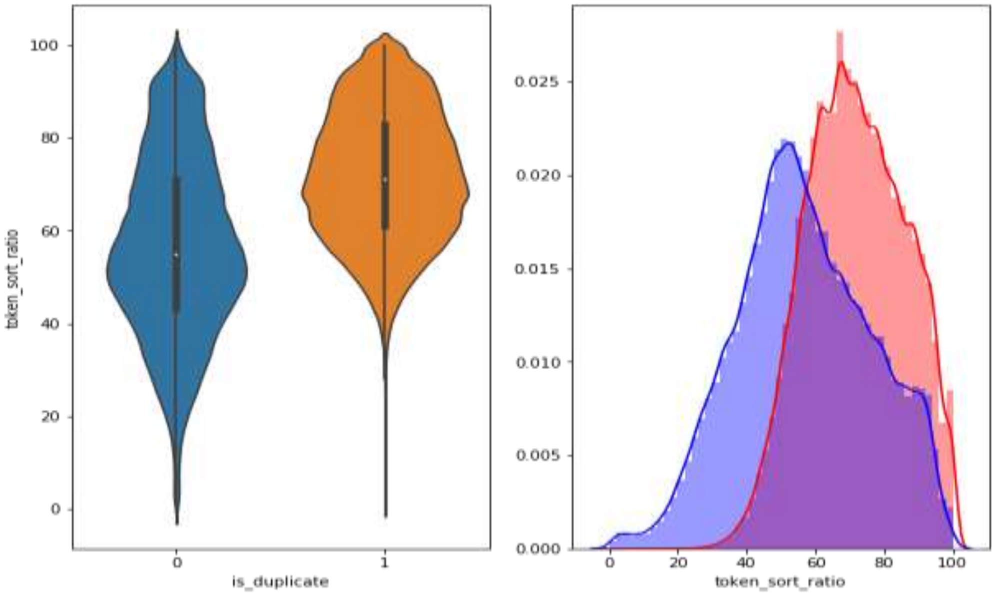

Figure 5: .Distribution of token ration of new and duplicate questions

The figure 6.Clearly shows that the proposed system deliberates the better result for prediction of new and duplicate questions.

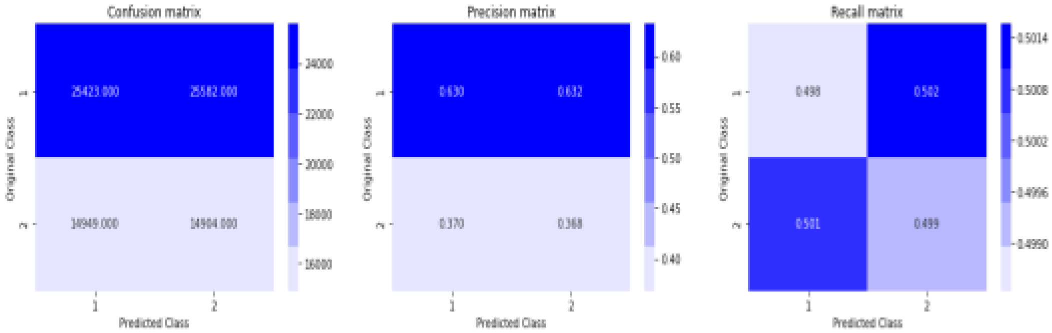

Figure 6: Calculation of confusion and prediction matrix for proposed system

V. CONCLUSION

The suggested techniques recognize the duplication questions effectively than the conventional techniques. First weights are assigned to all the questions posted in the forum and then uniqueness and repetition of words in queries as well as the duplication in query is analyzed. Finally normalization of vector is used to find the semantic analysis of queries also best classification method for semantic prediction. The experimental results proved that it

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

provides the better classification than competitive techniques. Weights are assigned separately for the new and duplicate questions. Classification of new and duplicate words is effectively found out using NLP techniques. The suggested techniques act as better tool for reducing the number of duplicated questions posted and time taken for the providing the response is also reduced.

REFERENCES

1. Alagarsamy, S., Kamatchi, K., Govindaraj, V., A Novel Technique for identification of tumor region in MR Brain Image,(pp-1061-1066), in proceedings of the IEEE Third International Conference on Electronics Communication and Aerospace Technology,2019.
2. Alagarsamy, S., Kamatchi, K., Govindaraj, V., Thiyagarajan, A., A fully automated hybrid methodology using Cuckoo-based fuzzy clustering technique for magnetic resonance brain image segmentation, Vol.27 (pp.317-332), International journal of Imaging systems and technology,2017.
3. Alagarsamy. S.,, Kamatchi, K., Govindaraj, V., Zhang, YD., Thiyagarajan, A., Multi-channeled MR brain image segmentation: A new automated approach combining BAT and clustering technique for better identification of heterogeneous tumors,Vol.39 (pp.1005-1035), Biocybernetics and Biomedical Engineering,2019.
4. Bowman, SR., Angeli, G., Potts, C., Manning, CD., A large annotated corpus for learning natural language inference, 2015.
5. Howland, P., Park, H., Generalizing discriminant analysis using the generalized singular value decomposition,Vol.26(pp. 995 – 1006), IEEE Transactions on Pattern Analysis and Machine Intelligence,2004.
6. Liu ,M., Lang, B., Zepeng, G., Zeeshan, A., Measuring similarity of academic articles with semantic profile and joint word embedding,Vol.22(pp. 619 – 632), Tsinghua Science and Technology, 2017.
7. Medjahed, B., Bouguettaya, A., A multilevel composability model for semantic Web services, Vol.(pp. 954 – 968), IEEE Transactions on Knowledge and Data Engineering, 2005.
8. Miller, A., WordNet: a lexical database for English, Vol.38 (pp.39-41),Communications of the ACM,1995.
9. Xie, X., Cai, X., Zhou,_J., Cao,_N., Wu, Y., A Semantic-Based Method for Visualizing Large Image Collections,Vol.25(pp. 2362 – 2377), IEEE Transactions on Visualization and Computer Graphics,2019.
10. Zhou, P., Shi, W., Tian, J., Qi ,Q, Li,B., Hao,H., Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification,(pp.2017-212), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016.
11. Prasanthi, E., & Deepa, N. (2019). Real time web based information using natural language processing (NLP) algorithm. Test Engineering and Management, 81(11-12), 5616-5620. Retrieved from www.scopus.com

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020