Soft Computing based Duplicate Text Identification in Online Community Websites

Basavesha D[#1], Dr. Y S Nijagunarya[*2]

[#]*Research Scholar, Department of Computer Science & Engg, Siddaganga Institute of Technology*
**Professor, Department of Computer Science & Engg, Siddaganga Institute of Technology*
Tumakuru, Karnataka, India

Abstract - *As the number of social media websites and applications are increasing the amount and the speed of data generation is also increasing and in turn the chances of having duplicates in the data are also increasing. The presence of duplicates will reduce the quality of data and also deteriorates the accuracy of the final results. Therefore, identifying and removing the duplicates is very important and it is considered to be a necessary step in data preprocessing and data integration. In this paper we have made an extensive review on the state-of-the art literature in the field of duplicate text identification. The paper consists of a survey on the works related to duplicate data identification, duplicate text identification and duplicate record identification. We have discussed generalized step by step procedure for duplicate text identification that is followed by most of the researchers. We described about word embedding techniques, similarity estimation techniques, and different soft computing techniques such as neural networks, fuzzy logic, evolutionary algorithms, Bayesian networks and support vector machines. We summarized the state-of-the-art works in three categories like, duplicate question identification in quora and stack overflow, text identification in documents and record identification in small and large datasets. Finally we also discussed about the different metrics used to measure the performance of the model developed for duplicate identification.*

Keywords — *Duplicate text, soft computing, neural network, fuzzy logic, bag-of-words.*

I. INTRODUCTION

Text mining is one of the important on-going research areas. The social media websites and online community cites are generating huge volume of data every day. The users are becoming the producers of this huge data both in the form of text and images. The rapid growth in volume of data makes it necessary to identify duplicate texts. Text analysis can be performed over online community databases for gathering preferences, for duplicate question detection, duplicate document detection, bug report detection etc. Duplicate text detection is very important task for data cleaning [1]. Data cleaning is a significant role in data mining. It is important to improve the quality of the data in data warehouse before applying data mining process. Data cleaning deals with identifying and removing the errors, missing values and removing the duplicates and inconsistencies to improve the quality of data. Removing duplicates in datasets actually means to remove the entities that are carrying same value for all the attributes [1]. Whereas removing duplicates in case of online community cites like quora and stack overflow actually means to identify the questions that are semantically same and can be answered with the same answer. By doing this it is possible to group all the questions together and provide answer that can satisfy all the questions. This ensures the quality and quantity of the content presented to the users. This enhances the user experience. Though many research works has been carried out on this area, still it is a challenging problem to detect duplicate text or record in quora, stack overflow, datasets and in other online community cites. The main reason behind this is the fact that, natural language is very expressive, same word gives different meaning based on situation and sequence, different words, phrases can be used to mean the same.

The step by step process for duplicate text detection is shown in the figure. The textual data is given as input to the word embedding system [2]. The word embedding system will represent the given input text in the form of vectors of real numbers. These vectors are then fed as input to the similarity checking techniques like Simhash and Minhash. The outcome of the similarity estimation is used to find the features of each text. The features are each text are computed together to find the distance between the both using some distance function. The distance can be computed either using Euclidian distance, Cosine distance or Manhattan distance.

During word embedding stage, the texts having same meaning will have same similar vector representation. Some of the most commonly used word embedding techniques are embedded layer, Word2Vec and Glove.

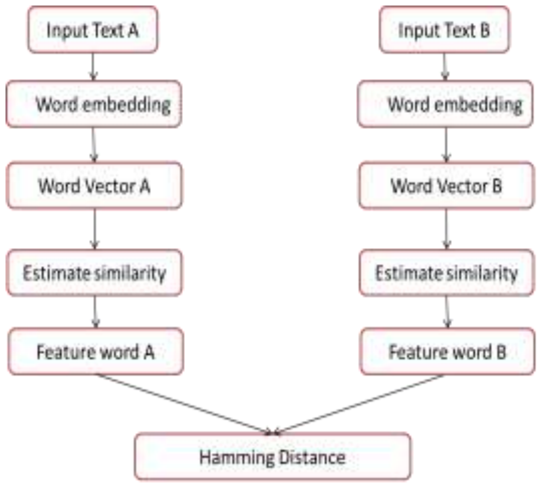

Figure: Flowchart for duplicate text detection

- **The embedding layer** is the one used with neural network models for natural language processing tasks. A clean text document is provided as input to the embedding layer [3]. Each word in the text is one-hot encoded such that the vector will have 50, 100 or 300 dimensions in the pre-specified vector space.
- **Word2Vec** is a statistical approach to provide word embedding for a textual dataset. Two methods can be used with Word2vec approach to learn word embedding. One method is the continuous bag-of-words model. In this case word embedding can be learnt by predicting the word based on the context. Another method is Continuous skip gram model which learns the word embedding by predicting the word based on the surrounding words
- **Glove (Global Vectors for Word Representation)** is an extension of Word2Vec approach. It is a combination of global statistical factorization technique like Latent Semantic Analysis and local statistical method Word2Vec. It is a regression model used for unsupervised learning of word representations.

The word vectors got from the word embedding system is fed as input to estimate similarity using the similarity checking techniques. Two most commonly used techniques to estimate similarity are Simhash and Minhash.

- **Simhash** is the technique used to detect the near duplicate texts. The texts are said to be similar if the hamming distance between them is as smaller as possible. The input text is divided into chunks; each chunk will have a hash function. The hash value of each chunk is represented as a vector with binary values. The bit values of the binary vectors are transformed into +1 and -1 based on the whether the bit value is 1 or 0
- **Minhash** is another technique to quickly estimate the similarity between the two texts of any kind of problem. In case of large scale clustering problems, Jaccard similarity is used to find the similarity between the two clusters.

After the similarity estimation is done, the features of the texts are then used to compute the hamming distance between the two texts. Hamming distance gives the number of bits different between the two feature vectors. If the dimensionality of the feature vector is high, the suitable method for computing the distance is Manhattan compared to Euclidean distance. Distance between two data points can be computed using cosine distance.

Soft computing is the technique used to study the science of reasoning, thinking and analyzing the real world problems. Duplicate data and text detection is carried out using different components of soft computing in state-of-the-art. Therefore in this work we are making an effort to give enough description about soft computing and its components.

The remaining segments of the paper are organized as follows: section II gives the detailed description about the soft computing and its techniques. Section III discusses the state-of-the-art techniques in the field of duplicate text detection. Section IV tells about the different performance metrics used to evaluate the duplicate detection model followed by conclusion.

II. SOFT COMPUTING AND ITS TECHNIQUES

Soft computing is a technique that provides imprecise result but still usable solution for complex computational problems. It is the fusion of methodologies that work systematically with flexible information processing capability to produces usable solutions to real world complex and ambiguous situations. The solutions obtained from soft computing are fuzzy in nature [4]. It differs from hard computing. Unlike hard computing, soft computing can tolerate, ambiguity, uncertainty, missing values, spelling errors and partial truth [5][6]. It is also an optimization technique that helps to make the solution better and better for the problems which are hard to solve.

In our research work we are developing a method for identifying the duplicate text in the given textual databases of online community cites. The entire research work includes step by step procedure like, studying the state-of-the-art in the area of soft computing; identifying and studying the feature extraction techniques used so far in the literature, the methods used for duplicate text identification and classification, the performance metrics used to evaluate the model's performance and then finally come out with an hybrid model which can perform better compared to state-of-the-art methods proposed so far.

In this work we are presenting an extensive survey of the works carried out in the literature. Here we present a review on various soft computing models available, different feature extraction techniques and performance evaluation techniques. The various soft computing models available for duplicate text identification are show in the figure below:

.

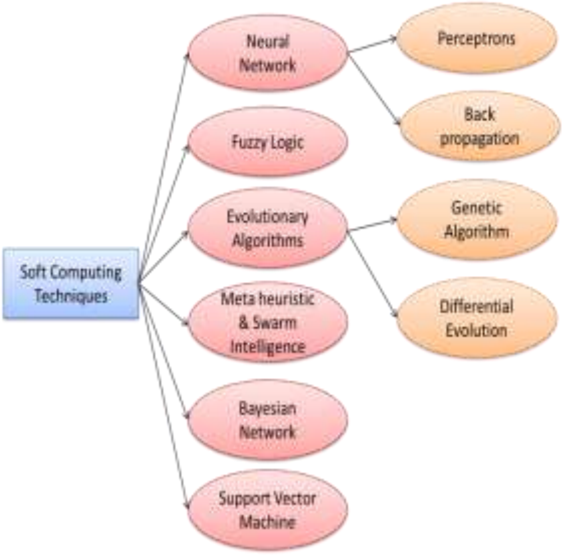

Figure: Soft computing models

A. Page Layout Neural Network

A neural network is a computational model that mimics the structure of human brain. Neural network is built with one input layer, one or more hidden layers and an output layer. Each layer is built with neurons. Neural network can be perceptron or back propagation. Perceptron neural network is a binary classifier which is linearly separable. Back propagation neural network works both in forward pass and backward pass. During backward pass the parameters like weights and biases of the layers are changed to reduce the error between predicted and actual output. There are many types of neural networks, feed forward neural network, recurrent neural network, convolutional neural network and kohonen self-organizing neural network. Many researchers have used neural network based models to perform text categorization, duplicate text identification and text mining. Yushi Homma et al [7] have used recurrent neural network and gated recurrent neural network to find duplicate questions in quora. Chakaveh Saedi et al [14] have used convolutional neural network to detect duplicated in quora. The results were compared with SVM, CNN showed good results compared to SVM.

B. Fuzzy Logic

Fuzzy logic is a many valued computing approach that produces the results between 0 and 1 instead of clear cut true and false value. It gives the degree of truthiness in the output produced. It is used to handle concepts with partial truth, the concepts whose degree of truthiness range between completely true and completely false. The architecture of fuzzy logic includes four parts, Rule Base, Fuzzification, Inference engine, De-fuzzification shown in the figure below. It is used in natural language processing and various applications in artificial intelligence. Dr. Murtadha M. Hamad et al [5] have worked on eliminating duplicates in data warehouse using fuzzy logic. They first identified the similarity using Q gram. A threshold of 68% was chosen based on the results obtained. The texts exceeding this threshold were then fed as input to fuzzy algorithm to determine whether the record is duplicate or not. The work was presented with an accuracy of 96%.

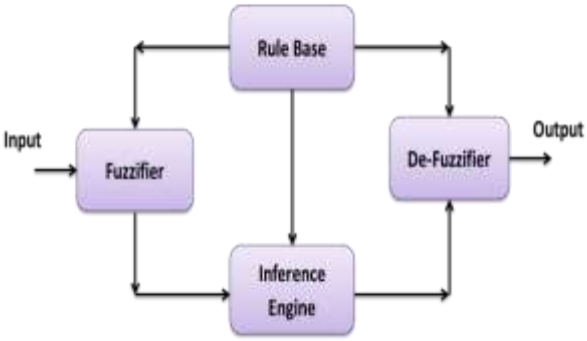

Figure: Fuzzy logic

C. Evolutionary Algorithms

These are heuristic based algorithms used to solve problems that take too long time to process exhaustively. Genetic algorithms and Differential evolution algorithms are the types of evolutionary algorithms. Genetic algorithms are the optimization techniques that work on the basis of genetics and natural selection. They are used to find optimal or near optimal solutions for the problems that would take lifetime to solve. Differential evolution optimizes the problem iteratively to improve the candidate solution. It is used to solve multidimensional real valued problems. It maintains a population of candidate solutions and generates new solutions by joining the existing ones. Finally the solution with best score is kept. Hamid Mohammadi et al [6] have proposed a new signature based approach to measure text similarity using cosine and genetic algorithms. They used large document collections such as CiteseerX, Enron etc. for the research work. They obtained around 87% precision, 98% recall and 92% f1_score with 10950ms of run time.

D. Bayesian Network

Bayesian network also called belief network are based on probability theory. They use probabilistic graphical model to identify the relationships between the attributes. They represent conditional dependence in the form of directed graphs. Nodes in the graph represent attributes and edges in the graph represent conditional dependence. There are many works in the state-of-the-art that have used Bayesian network to predict duplicate text in the corpus. Nithya. P et al [8] have worked on identifying duplicate text in XML data. They used Bayesian network to find the probabilities of the XML elements. They also used decision tree induction pruning strategy to improve the network efficiency. Nikhil Gawande et al [9] also worked on detecting duplicates in hierarchical data using Bayesian network. They proposed a novel XMLDup method based on Bayesian network to compute the probabilities of two XML nodes not merely on the basis of node value but also the structure of the node. They used real time restaurant data set for the experiment and came out with an impressive precision, recall and effectiveness in the experiment..

E. Support Vector Machine

Support Vector Machine is a supervised learning technique used for both classification and regression problems. It performs both linear and non-linear classification. For non-linear classification it uses kernel trick to transform the dimensions of the data. Based on the transformations made, it computes an optimal boundary to classify the data objects. Girija M [10] alone has worked on duplicate data detection on multiple web databases using support vector machine. The author used unsupervised duplicate detection algorithm to compute the similarity vectors of the selected dataset. Further support vector machine is used for classifying the data. The experiment was divided into two categories; one querying having random words and the other analysis of restaurant dataset. The author was successfully able to identify 112 duplicate pairs.

III.RELATED WORK`

Duplicate detection is not a completely new problem. Many research works have been carried out with different approaches to develop a method for duplicate discovery from decades. Some of the state-of-the-art works in this field are summarized in this section.

Yushi Homma et al [7] have worked on determining the semantic equivalence between the pairs of questions in quora dataset. They used deep learning based Siamese gated recurrent unit neural network for encoding each input sentence. They tried different distance measures to predict equivalence of the sentences based on vector outputs of the neural network. Two questions are said to be semantically equivalent if they cab ne answered exactly by the same answer. They used completely labeled dataset from quora. They started the experiment with data pre-processing in which they used Stanford Tokenizer from standard Stanford Core NLP suite. They performed spellcheck pre-processing followed by data augmentation, hyper parameter search. In this project, two types of neural network were used to encode the each input sentence; recurrent neural network and gated recurrent unit. Both the neural network outputs a sentence vector of dimension H, nothing but the hidden vector size in the neural network. After each pair of the sentence is encoded, the distance between the two is calculated using 3 distance measures, cosine, Euclidean and also Manhattan. Even after trying three measures, finally they considered distance measure by calculating using neural network with softmax classifier. The prediction whether the sentence is duplicate or not was done using logistic regression. They obtained prediction accuracy of about 0.8627 and f1_score of about 0.8105.

Travis Addair [11] has worked on determining whether two questions are asking for the same answer, which indirectly means that whether the two questions are similar to each other. The author has come out with a series of models using deep learning approach. The models include convolutional neural network, long short term memory neural network and a hybrid model. These models are built on top of Siamese network architecture and multilayer perceptron. All three models gave outstanding performance when compared to traditional natural language processing techniques. In this work each question is considered as one dimensional vector. These vectors are converted to pre-trained word embeddings with Glove in the embedding layer. The outcome of embedding layer is passed on to the encoding layer which will output the one dimensional feature vector. The output feature vectors are then combined and passed on to multilayer perceptron. At last MLP produces the final output. For the encoding layer three models one after the other were used. They obtained accuracy of 0.8027, 0.8107 and 0.8105 for CNN, LSTM and hybrid model respectively. Similarly, an f1_score of 0.7223, 0.7570 and 0.7466 for CNN, LSTM and hybrid model respectively.

Lei Guo et al [12], have worked on duplicate question detection using Quora datasets. They first vectorized the questions, extracted features, trained and then used machine learning techniques for prediction. The prediction is done based on vectors and the features extracted. They used two different approaches to detect the duplicates with different vectorization methods and feature extraction methods. The two approaches are; first approach with Word2Vec and TF-IDF score, second approach was using neural network with term frequency. They used different methods for classification, KNN, SVM and Random forest and obtained an accuracy of 80% in both the approaches.

Sujith Viswanathan et al [13] worked on detecting duplicates using Quora and Twitter datasets. Detecting duplicates helps for deduplication, a process of removing duplicates to improve the data quality. The authors say that detecting duplicates using natural language processing method is less accurate. Hence they used machine learning techniques to improve the accuracy. They used six supervised ML techniques to perform classification between duplicates and non-duplicate sentences. They used word share and TF-IDF word share as features to identify duplicates. TF-IDF is an important feature measure to identify duplicates. The machine learning algorithms used for classification include, Logistic Regression, Decision Tree, SVM, KNN, Naïve Bayes and Random Forest. The experiment performed well with the accuracy of 78.6, 70.3, 66.0, 78.0, 66.1 and 78.1 respectively for all the ML techniques listed above respectively.

Chakaveh Saedi et al [14] have made a contribution in developing automatic duplicate question detection in large corpus datasets of online community cite like quora. The authors say that the performance of the duplicate question detector system does not depend on the grammatical issues in the text or the questions in the community forums. The performance will not even depend on the lengths of the questions quoted instead, the performance mainly influenced by the size of the datasets used. The performance is degraded when the system is trained with a high volume data from many different sources and different domains. Performance also degrades as we move from more narrow domains to generic domain. The authors used rule based jaccard index to measure the similarity between the two questions. They also used support vector machine and deep convolutional neural network to illustrate the performance of duplicate question detection with increasing size of the input dataset. The rule based approach gave a duplicate detection accuracy of 69% for the input dataset consisting of 7k pairs of questions and 69.50% for 300k question pairs. SVM was able to give accuracy of 67.64 % for 7k input question pairs and 66.55% for 165k pairs. Finally deep convolutional neural network gave impressive results with accuracy of 62.29% for 7k input question pairs and gradually increased the accuracy with the increase in the size of the input dataset. It reached to 77.64% for the input dataset of 300k question pairs.

Jin Gao et al [15] have worked on duplicate text detection in short texts using bag-of-words algorithm. Words in the text are represented in the form of vectors. These vectors are provided as input elements to Simhash algorithm which in turn outputs the input vector in the form 64 bits sequences. These sequences are compared using hamming distances. The results are obtained comparing with the pre-set threshold value. The authors have also worked on improving the results by incorporating the weight concept. They compared the results obtained with the unweighted Word2Vec and TF-IDF methods. For the experiment the authors used dataset obtained from the sick corpus. They illustrated that weighted Word2Vec performed with good results compared to unweighted. Weighted Word2Vec gave 68.6% accuracy and 43.6% of fl_score value. Unweighted Word2Vec gave 68.5 and 39.4% of accuracy and f1_score respectively. Unweighted TF-IDF gave 65.9 and 21.2% of accuracy and f1_score respectively.

Yifang Sun et al [16] have proposed a method for near duplicate text detection using signatures. They used collection frequency of Q-grams in the proposed work and compared their method with winnowing which is also a signature selection algorithm. They proposed a novel concept called k-stability and applied this concept with all winnowing algorithms. They also proposed another variant model with winnowing algorithm, named as frequency biased winnowing. This approach achieved good accuracy and efficiency compared to other similar works in the literature. For the experiment the authors used PAN-PC-10 dataset available publicly. Frequency biased winnowing approach gave impressive f1_score of 77.5% with q=4 while normal winnowing achieved only 49.6% with q=50 and 74.5 with q=10.

IV. PERFORMANCE METRICS

As we see in most of the related works in state-of-the-art, the commonly used metrics to evaluate the performance of the model in identifying duplicate text are, accuracy, precision, recall and f1_score [17][18]. Accuracy is the percentage of correct predictions over total number of predictions or the number of objects in the test dataset. The general formula for calculating accuracy is as shown below:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ data\ examples}$$

Precision is the measure of exactness while recall is the measure of completeness. High value of precision indicates that the model returned more relevant outputs than the irrelevant output [19][20][21]. High value of recall indicates that the model has returned more of relevant output only. Precision and recall can be obtained with the following formulas respectively:

$$Precision = \frac{relevant\ examples + retrieved\ examples}{retrieved\ examples}$$

$$Recall = \frac{relevant\ examples + retrieved\ examples}{relevant\ examples}$$

F1_score is the measure of test accuracy. It is computed using precision and recall values. The best value of f1_score is 1. It is most often used for measuring the performance of search, query classification and document classification [22][23][24]. For the case of binary classification, it is less informative compared to Mathew correlation coefficient. It can be computed using the formula shown below giving equal importance to both precision and recall [25]:

$$f1\_score = 2 * \frac{precision * recall}{precision + recall}$$

V. CONCLUSION

Duplicate detection is one of the crucial tasks. In case of online community cites like quora and stack overflow, if the duplicate questions are identified, they can be grouped together and answers from different experts can be made available to all those duplicate questions in one hit. In case of small and large datasets used for research works, the presence of duplicate records will reduce the accuracy of the output results. Therefore it is very important to identify and remove the duplicate data and improve the quality of the documents, website data and datasets. In this paper we have made a review on state-of-the-art literature in the field of duplicate text identification. Our work gives insights about different techniques used for the duplicate identification for the upcoming researchers in this area.

REFERENCES

[1] E V Sharapova and R V Sharapov, "*The problem of fuzzy duplicate detection of large texts*", IV International Conference on "Information Technology and Nanotechnology" (ITNT-2018).

[2] John Rathbone*, Matt Carter, Tammy Hoffmann and Paul Glasziou, "*Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module*", Rathbone et al. Systematic Reviews 2014, 4:6.

[3] Thorsten Papenbrock, Arvid Heise, and Felix Naumann, "*Progressive Duplicate Detection*", IEEE Transactions on Knowledge and Data Engineering, 1041-4347 (c) 2013 IEEE.

[4] Yun Zhang, David Lo, Xin Xia, Jian-Ling Sun, "*Multi-Factor Duplicate Question Detection in Stack Overflow*", JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(5): 981–997 Sept. 2015.

[5] John Rathbone*, Matt Carter, Tammy Hoffmann and Paul Glasziou, "*Better duplicate detection for systematic reviewers: evaluation of Systematic Review Assistant-Deduplication Module*", Rathbone et al. Systematic Reviews 2014, 4:6.

[6] Thorsten Papenbrock, Arvid Heise, and Felix Naumann, "*Progressive Duplicate Detection", IEEE Transactions on Knowledge and Data Engineering, 1041-4347 (c) 2013 IEEE.

[7] Yushi Homma, Stuart Sy, Christopher Yeh, "*Detecting Duplicate Questions with Deep Learning*", 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain

[8] Nithya. P, Vinothini. K, "*Duplicate Detection in XML Data Using Probabilistic Duplicate Detection Algortihm*", International Journal of Engineering Research & Technology, Vol. 3 Issue 1, January – 2014.

[9] Nikhil Gawande, S. R. Todamal, " *A Survey on Duplicate Detection in Hierarchical Data*", International Journal of Science and Research, Volume 3 Issue 12, December 2014.

[10] Ms. Girija. M, "*Handling Duplicate Data Detection Of Query Result From Multiple Web Databases Using Unsupervised Duplicate Detection With Blocking Algorithm*", International Research Journal of Engineering and Technology, Volume: 03 Issue: 04 | Apr-2016

[11] Travis Addair, "Duplicate questin pair detection with deep learning". https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2759336.pdf

[12] Lei Guo, Chong Li, Haiming Tian, "*Duplicate Quora Questions Detection*", https://pdfs.semanticscholar.org/4c19/2b8f45b1e913ee7da32624cd7559eccb0890.pdf

[13] Sujith Viswanathan, Nikhil Damodaran, Anson Simon, Anon George, M. Anand Kumar and K. P. Soman, "*Detection of Duplicates in Quora and Twitter Corpus*", J. D. Peter et al. (eds.), Advances in Big Data and Cloud Computing, Advances in Intelligent Systems and Computing 750, © Springer Nature Singapore Pte Ltd. 2019.

[14] Chakaveh Saedi, Jo˜ao Rodrigues, Jo˜ao Silva, Ant´onio Branco, Vladislav Maraev, "*Learning Profiles in Duplicate Question Detection*", IEEE International Conference on Information Reuse and Integration (IRI), 2017.

[15] Jin Gao, Yahao He, Xiaoyan Zhang, Yamei Xia, "*Duplicate Short text detection using Word2Vec*", 978-1-5386-0497-7/17/$31.00 ©2017 IEEE.

[16] Yifang Sun, Jianbin Qin, and Wei Wang, "*Near Duplicate Text Detection Using Frequency-Biased Signatures*", X. Lin et al. (Eds.): WISE 2013, Part I, LNCS 8180, pp. 277–291, 2013.c_Springer-Verlag Berlin Heidelberg 2013.

[17] Jo˜ao Rodrigues, Chakaveh Saedi, Ant´onio Branco and Jo˜ao Silva, "*Semantic Equivalence Detection: Are Interrogatives Harder than Declaratives?".* http://www.di.fc.ul.pt/~ahb/pubs/2018RodriguesSaediBrancoEtAl.pdf.

[18] Zainab Imtiaz, Muhammad Umer, Muhammad Ahmad, Saleem Ullah, Gyu Sang Choi, and Arif Mehmood, "*Duplicate Questions Pair Detection Using Siamese MaLSTM*", IEEE Access, VOLUME , 2019.

[19] Marios Poulos, "Near Duplicate Text Detection using Graph Depiction". https://www.researchgate.net/publication/311756563_Near_duplicate_text_detection_using_graph_depiction.

[20] Abram Hindle1 · Anahita Alipour1 · Eleni Stroulia, "*A contextual approach towards more accurate duplicate bug report detection and ranking*", Springer Science+Business Media New York 2015.

[21] JIANKUN YU, MENGRONG LI, DENGYIN ZHANG, "*Duplicate text detection based on LCS algorithm*", 2nd Information Technology and Mechatronics Engineering Conference (ITOEC 2016).

[22] P.Lakshmi Prasanna, S.Manogni , P.Tejaswini , K.Tanmay Kumar , K.Manasa, "*Document Classification Using KNN with Fuzzy Bags of Word Representation*", International Journal of Recent Technology and Engineering ISSN: 2277-3878, Volume-7, Issue-6S, March 2019.

[23] Yuliang Xiu, Xiaoting Jiang, Weiyu Cheng, Bowen Zhang, "Quora Question Pairs @ Kaggle", Shanghai Jiao Tong University, X033525, June 7, 2017.

[24] Ramya R S, Venugopal K R, Iyengar S S & Patnaik L, "*Feature Extraction and Duplicate Detection for Text Mining: A Survey*", Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 16 Issue 5 Version 1.0 Year 2016.

[25] Nayana, Y., J. Gopinath, and L. Girish. "*DDoS mitigation using Software Defined Network.*" International Journal of Engineering Trends and Technology (IJETT) 24.5 (2015): 258-264

Table: Summary of the related work

	Author	**Algorithm**	**Dataset**	**Accuracy**	**F1_score**
Duplicate question detection	Yushi Homma et al [7]	recurrent neural network and gated recurrent unit	Quora	0.8627	0.8105
	Travis Addair [11]	CNN, LSTM and hybrid model	Quora	0.8027, 0.8107 and 0.8105	0.7223, 0.7570 and 0.7466
	Lei Guo et al [12]	1)Word2Vec and TF-IDF 2)Neural network	Quora	80%	--
	Sujith Viswanathan et al [13]	Logistic Regression, Decision Tree, SVM, KNN, Naïve Bayes and Random Forest	Quora & Twitter	78.6, 70.3, 66.0, 78.0, 66.1 78.1	0.78 0.70 0.67 0.77 0.74 0.77
	Chakaveh Saedi et al [14]	Rule based Jaccard index SVM DCNN	Quora	69.50 66.55 77.64	--
Duplicate text detection	Jin Gao et al [15]	Weighted Word2Vec Unweighted Word2Vec TF-IDF	SICK Corpus	68.6 68.5 65.9	43.6 39.4 21.2
	Yifang Sun et al [16]	Frequency biased winnowing Winnowing	PAN-PC-10	--	77.5 (q=4) 49.6 (q=50) 74.5 (q=10)
	E V Sharapova et al [1]	Fuzzy search algorithm	Large text document	--	Recall = 99%
	Hamid Mohammadi et al [6]	Cosine and Genetic algorithm	CiteseerX, Enron, Gold Set of Near-duplicate News Articles	--	92%
Duplicate record detection in Large datasets or data warehouse	Dr. Murtadha M. Hamad et al [5]	Fuzzy logic (FL) and Q-gram	Data warehouse	96%	--