Turkish Journal of Computer and Mathematics Education *Vol.12 No.6 (2021), 4024-4033*
Research Article

Deep Neural Model for Duplicate Question Detection Using Support Vector Machines (Svm)

Seema Rani[a], AvadheshKumar[b], Naresh Kumar[c], Sanjay Kumar[d]

[a]Research Scholar, School of Computing Science & Engineering., Galgotias University, Greater Noida, Uttar Pradesh, India
seemanandal011@gmail.com .

[b]Professor, School of Computing Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
avadheshkumar@galgotiasuniversity.edu.in

[c]Professor, School of Computing Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India
naresh.dhull@gmail.com

[d]Assistant Professor, Information Technology Department, Galgotias College of Engineering & Technology, Gr. Noida, India
skhakhil@gmail.com

ABSTRACT

Stack Overflow has been a core component of the toolset of the developer. This rise in influence was followed by a Stack Overflow group initiative to preserve the Content consistency. One of the threats that threaten the persistent growth of duplicated questions is consistency. Resolution of Prior research on the automated identification of this problem Questions multiplied. DupPredictorand Dupe are two essential solutions. We carried out aDupPredictor and Dupe's observational replica analysis. While the findings are important, both works are not freely accessible, so hindering their introduction. The following work depends on them in the science literature. Wecarried out a DupPredictor observational replica analysis And Dupe. Our findings are not stable in various ways,sets of methods and data sets. To illustrate the replication barriersfor these methods and approaches are high. In addition, if more is necessary we observe a decrease in efficiency of our two recentlyRecall-rate reproductions over time, as the numberof questions is increased. The following are our resultsof study on the identification of questions duplicating inquiry to claim and Respond Communities with their assumptions. The findings of this paper are systemic and comparative tests with main technique stylesfor predictive question identification duplication detectionApplied to increasingly broad data collections, such thatto research the profiles of learning of this mission, approaches and assesses the merits. This research has been carried out by using the latest publication for research purposes, Probable a new engine by Quora online reply query dataset with more than 100000 marked pairs Duplicate portions of questioning are components.

Keywords:Quora, Duplicate question, Deep Neuralmodel, speech processing, Community-based question answering

1. INTRODUCTION

The nature of the teaching and learning ecosystem modified by Internet connectivity and the omnipresent network. With the recognized need for peer feedback and the trust in social networking material online, people are looking for decision making information. Anyone out of there, the Wiki Answers, Friend feed and Stack Overflow explains the boom in content and social engagement in the common knowledge base. The exceptional amounts are found in the millions of users and thousands of questions posed and replied every day to probing responses (Q&A). Direct consumers, configure preferences and offer choices based on interests [1].While these Q & A channels allow for instant details, the response time is high and the accuracy of the response is underminedwith the inflow of

questions and answers. Moreover, duplicating material corrupts the process for filtering. The emphasis must therefore be changed from "informative excess" to "filter failure" hitches. Therefore, clever, smart and semanthrough filters are now essential to help direct consumers, configure tastes and deliver interest-based choices.

Semantic equivalence is a permanent challenge in the processing of natural languages [2]. Duplicate question identification attempts to align question pairs semantically to group together related intentions. The semantically question coupled problem is defined formally as: in the event of a question pair q1 and q2, create a model that can be graded, as stated in (1):

$$C(q1, q2) \rightarrow 0 \text{ or } 1 \quad (1)$$

Where, 1 represents semantic equivalence of q1 and q2 and are duplicates and 0 represents pair is non-duplicate.

2. LITERATURE SURVEY

Transparency in sharp and learning, the use of the traditional Q&A platform was distinguished by accessing alternative viewpoints, communication and engagement**[3]**. In the other hand, it is difficult and long to filter out the appropriate details for improved user experience (best answers/semantically matched questions/experts). Unsafe user choice, confusion regarding the fluidity and duplicate of the questions as well as the imprecision associated with the broad and varied responses and user base are some of the issues associated with them that hinder better information filtering platforms.Detecting synonymy among sentences is a tenacious problem in natural language processing.

Dey et al. [3] proved the ability to detect semantically similar blogs using handcrafted set of features on the SemEval-2015 dataset by Support Vector Machines (SVMs). In recent years, profound learning strategies have made considerable strides. They are used to detect semantically similar sentences with a Siamese nerve network architrave**[3].** Siamese neural network model uses the same neural network to encode two phrases individually**[5].**Detecting duplicate questions efficiently not only saves seekers time to find the correct answer, it also reduces the initiative of writers to answer several iterations of the same question. Studies on the duplicate Q&A discovery were carried out after the competition was launched by Kaggle which called on participants to recognize double questions. **[7]** Quora has carried out a comprehensive study of the influence of three separate link networks based on user theme index, social relation maps and a map linking questions. Quora was analyzed in depth.

Saedi et al. [8] submitted an exhaustive analytical review of the Quora dataset for automated duplicate query identification. The findings of structural experiments of duplicate questions identification in some existing methods and new techniques were made known by **Rodrigues et al. [9]** in their work.

In order to resolve Quoran duplication identification, **Chen et al [11]** employed basic functional engineering as well as more convoluted neural-network models.

DupPredictoron Stack Overload has been suggested by **Zhang et al. [11],** which can classify possible new problems by considering several factors considering its name, definition, question-related tags and latent topics.

In an overview of the grounds behind redundant questions, **Ahasanuzzaman etal. [12]**Suggested the model called Dupe based on a classification technology based on logistic regression.

In a reproduction of DupPredictor and Dup called DupPredictorRep and DupeRep respectively, **Silva et al. [13]** carried out an observational review. In 2019, authors **Viswanathan et al. [14]** recorded the identification of duplicates in Quora and Twitter Corpusing machine learning techniques such as random forests, logistic regression, vector support machine (SVM) and decision tree, using terms and TF-IDF to calculate similarity, for the defined classification of sentences in paraphrases or non-paraphrases.

In order to meet semantically equal questions in Quora, **Kaur and Gulati[15]** suggested an algorithm for graphic centred matching. More recently, **Shirani et al. [16]** have built a major StackOverflow dataset with a question-question relationship of over 250K pairs.The work on repeat query pairing is clearly primarily confined only to English as the existing Q&A sites impart monolingual assistance (only in English).

Chandu et al. [17] introduced an Indian language code-mixed Factoid Q & A system, Hindi and Tamil mixed together with English. As far as our understanding is concerned, no study in the literature has been recorded in the area "multiple language or two-lingual similarity pairs." Main and secondary experiments deal mainly with questions in English and discard questions as noise in transliterated or in other languages. The latest authors' study **[18]**discusses a cross-language cQA query search by machine which translates the inquiries into our target language and continues with the monolingual question search model.

3. DATASET

We define the selection and exploratory details in this section.Data processing, presentation of data and the method of data cleaning.

3.1 Data Collection

Data from the 1st Quora Dataset for this study work amazons S32 release hosted. The cumulative number of rows is 35840 the dataset showing that the cumulative query is 34560.The cumulative file size of pairs is 48.3 MB.

For word embedding, pre-trained GloVe word vectors are used.SNLI project site provides **GloVe [19]** pre-trained vectorsGlove. - Glove. We used to translate words to vectors to measure distance**[20]** Google News-vectors-negative300.bin.gz, with Google News VectorsThree million vocabulary and three hundred dimensions.

3.2 Data Exploration

We have conducted the necessary data set statistics to supportdetailed Quora duplicate question dataset understanding. The dataset includes six attributes. Everyone the features of the columns are meaningful this lines. The row. As described below, the definition of columns is Table 2.

Table1. Dataset column overview

Column Name	Description
id	A single ID allocated for each row Data collection. Data set. The first row has an identifier of 0, and ID 358724 in the last row
qid1	A special id for the topic at issue 1 Column.
qid2	A single ID for the topic in question2 Column.
question1	The real question to be answered in question 1 D to question2 Compare
question2	The exact issue of question2 is includedD to question2 Compare
is_duplicate	The parasite of question pair. -ve Of question pair. 0 means incorrect i.e. Is not duplicate pair. 1 shows real query pairi.e. duplicate query pair.

Duplicates and negative samples are semantically not duplicatePeer.

Table 2: Distribution of class marks

Sample Positive (1)	134256
Sample Negative (0)	240520
Issue Complete Pairs	403562

The x-axis represents the number in the histogram Fig.1 questions of occasions appear, and y-axis or bar height this is how many other questions exist for the number of occurrencesin data collection. In data set. The bulk can be seen in the graph the first bar shows the rare incident, the second bar shows the presence number less than 50 times. Twice and so forth, of question.

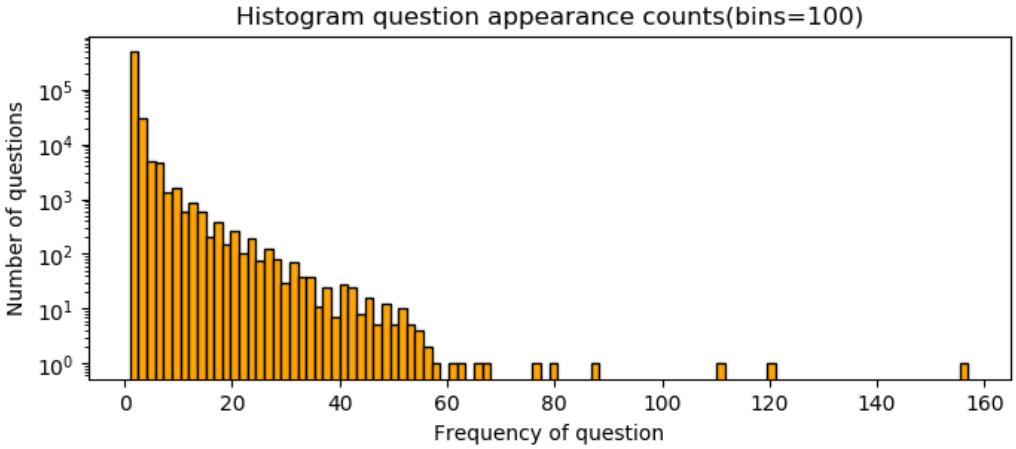

Figure 1: Distribution of the data in question [21]

4. BACKGROUND

This segment includes a short summary of the raw features dataset, different computer and deep neural learningLayers of experimental use.

4.1 Feature Engineering

We fell ID, qid1, and qid2 from the first three columnsOriginal data set and generated more helpful functions, and then we have two question1, question2 columns and are duplicate class name.So initially we have got total 27 new derived attributes causing 27 columns in the dataset supplied as a computer inputClassifiers for learning.

Set 1 Original Feature

(1) Question 1: Data in the question1 topic ofdataset.

(2) Question 2: Question 2, in the column Question 2of dataset.

(3) isduplicate: the class symbol is 1 for similar intent and 0fordistinct.

Set 2 Basic Features

(4) Question Length 1: Question Length 1 contains Characters, white spaces and punctuation.

(5) Question length 2: Question length 2 contains Characters, white spaces and punctuation.

(6) Difference in the duration of the questions:Question1 and Question2 corresponding length.

(7) Character number in q1: number of characters, here blankspaces in sentence are omitted.

(8) Character number in quesion2: Separate character numberexcept blank spaces in question2.

(9) Word count of quesion1: number of words in issue1, all words used.

(10) Quesion2 word number: number of words in question2 including number of repeated words.

(11) Amount of words in q1 and q2: different generic words Question1 and Question2 equivalent words.

Set 3 Fuzzy Functions

(12) Fact: Quick ratio comparison of the Qratio:The size of two query strings is between 0 and 100. Similarly questions are more valuable.

(13) Wratio: The weighted ratio is the value usedvarious algorithms for the score and returns estimation two question strings' best ratio. The number of scores is between 0 and 100.

(14) Figure set ratio: Figure set of token [23] is computed Lines are divided into three parts by the strings. Piece one popular string is arranged as an intersection, common strings

As sorted remains, the other pieces of each question. It then measure comparable results with each sorted intersection the variation of the triple crossing and the triple remnants Strict. The score range is between 0 and 100.

(15) Token type ratio: The sorting token of the strings Select the strings and rejoin into strings alphabetically. It the transformed strings are then compared by the return score ratios.

5. PROPOSED METHODOLOGY

The dataset is collected to delete 100 query pairs from different social media sites, including Quora and Trip Advisor. There are query pairs where one topic is in English, the other in Hinglish. Duplicate and non duplicate marking is annotated on the data collection. A subset of the dataset is shown in table 1 below.

1.Term Still WORD

Output is not recognized as the Deep Models. Voice or email to make the details comprehensible each query needs to be victories for certain models. Inside us. Model suggested, the upper layer contains the layer appropriate Question pairs as feedback and each word is translated to a Vector. The integration dimension is 400 and the limit 20 is the length of the sequence. Three related works in this Google News Vector word are embedding, Fast Text creeping, and sing the sub word.

2. GoogleNewsVector

Google offers news-based pre-trained term insertion Body. Includes three million English in this phrase 200 dimensional terms, offering three billion word vectors [24].

3.Text Fast

Fast Text is a library of studying language representation Facilitated by the study team on Face book. It requires 2 million famous 300 dimensional crawl words, supplying the word-vectors are 600 billion. That's better from Google Term embedding since the character level of n-gram is given Word representation [24].

4. SUBWORD Fast Text

Fast Text Sub word comprises 2 million qualified word vectors with Popular Crawl sub word details (600B tokens). Sub word incorporation brings us more information by conversion

In its sub words, any word. If the words are to be sent the corresponding sub words are, of word 'where' with n D 3, 'Her' and 'ere.' 'Whe' and 'her.' Finally, the dictionary is given these sub words are united [25].

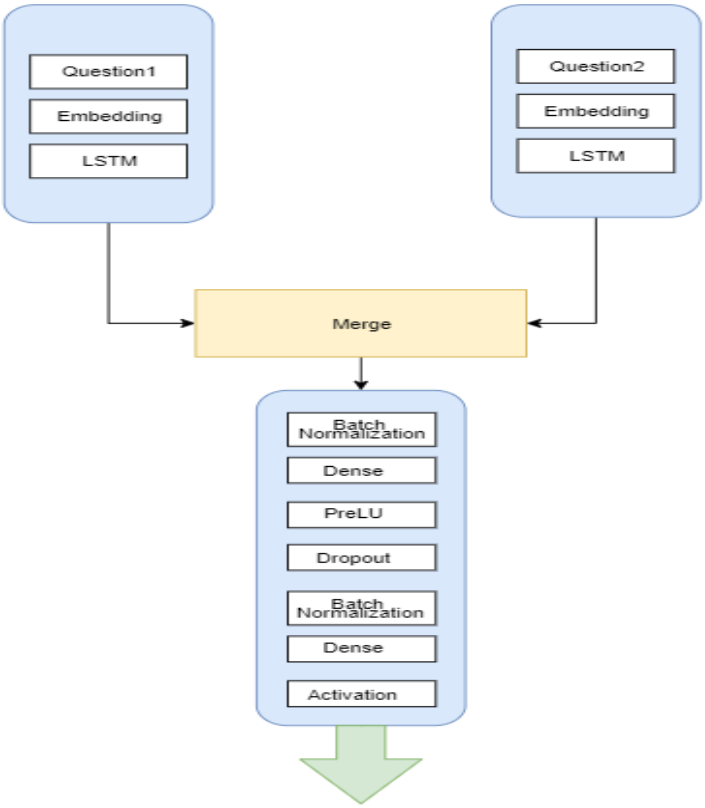

Figure 2: Clear Architecture of the Neural Network Two entries [19].

6.RESULTS AND DISCUSSION

This section addresses assessment measures and benchmarking the consequence.

6.1 Metrics of Assessment

The measurement collection is the most critical step in the assessmenthow we calculate the efficiency of our models howour paradigm and the baselines against each other.

Accuracy:Precision is the ratio of the right total numbermodels estimate the total number of predictionsthe model was asked.

Know: Remember or sensitivity is the expected positive ratio. Samples which are positive for the overall actual numbercomplete optimistic survey forecasts.

6.2 Classifying Baseline Model

We have educated our model and then assessed our test forecast.Data for the baseline for our computer algorithmsUsed in the inquiry. The accuracy and F score of Table 3 is showing simple learning frameworks for our machine.

Table 3: The baseline results on the basis of the 25 characteristics of a comparison data set of the standard machine learning classifier

Classifiers	Account	F--Score
Next neighbours	0.6275	0.6031
Adapter	0.6041	0.7936
Stepping Gradient	0.6417	0.6326
Tree for Decision	0.6271	0.6176
Forest Random	0.6054	0.7992
Extra trees	0.6039	0.6016

As shown in Table 3, the Xgboost model is obvious overtakes the accuracy of all other selected classifiersScore between 0.6416 and 0.6326 for F.

6.3 Study of Function Value

All seven computers have evaluated the function worthLearning classifiers used in tests and performedExperiments. We have chosen based on our characteristic qualitiesOut of 20 derived functions, the top 15 features.

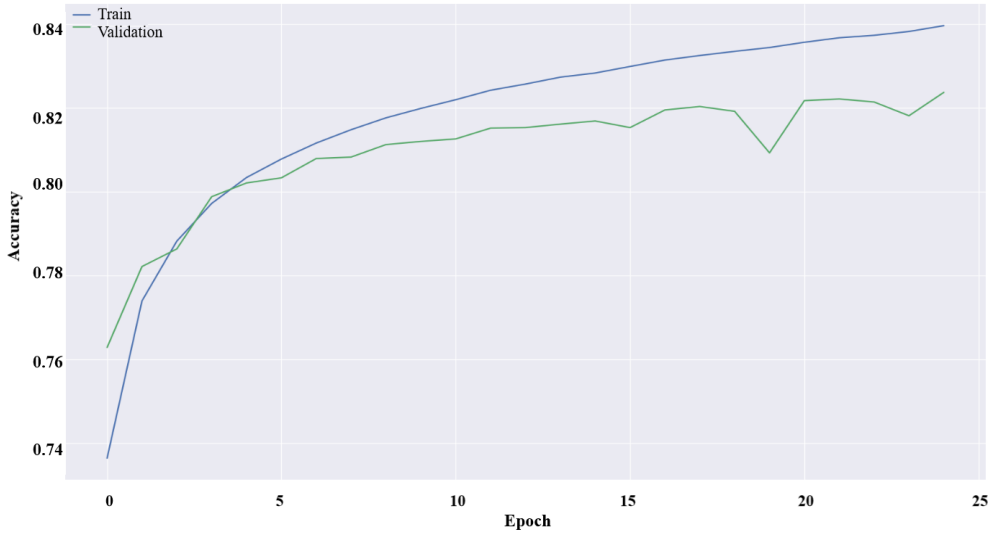

Figure 3: ML classifier performance and Accuracy

6.4 Performance of DQDModel

The Manhattan hybrid Siamese neural network model, which takes a query pair to be an input, has accomplished a similarity matching task and determines whether the questions in the pair are duplicated or not using the MLP classification. The consistency and F-score efficiency has been tested. 'Exactness is characterized as proximity to the true value of the calculation, i.e. as a proportion of true positive and true negative values between total cases inspected'. "The harmonic mean of precision and recall is measured in F-score". All values are displayed in proportions. The chart belowshowsDQDHinglish model's accuracy and F-score.

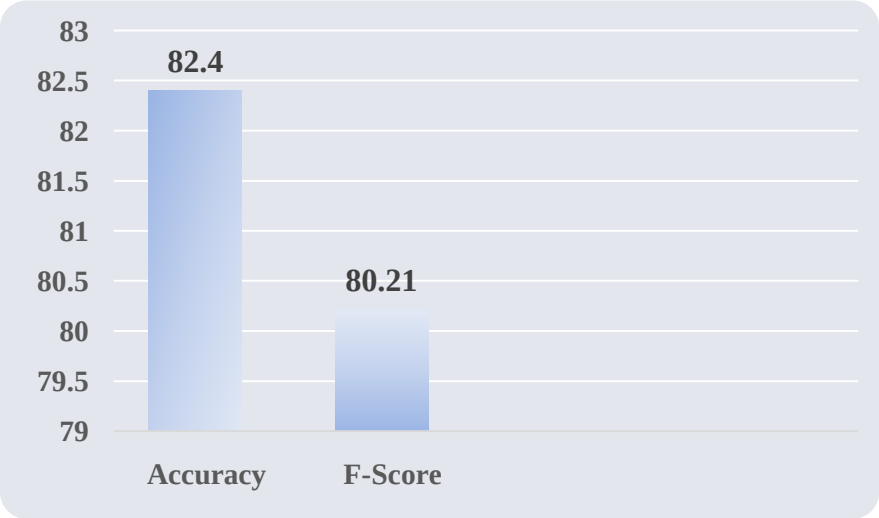

Figure 4:Accracy and F-Scorue

Minimum after the function fell. The comparable figures show Figures 3 Accuracy and F1 score visualization before and after the functionDrop**.**

7.CONCLUSIONS

During the trial, we are making sure the train and test details are separated into 75/18. We also make sure that the class labels are assigned accordingly in the test data set as in our original dataset. Both hyper parameters are chosen based on grid searches conducted on a 13 percent data collection, so that we do not overwrite our result. Our findings are very close to the state of the art precision of Quora of 89%. Quora has used its own term embedding from the Quora Corpus data collection, which is very unique to Quota query format, etc. Which is a big explanation for the disparity in outcomes? Although the general embedding of Glove is used, our findings are approaches which are more applicable to both general questions and answers.

REFERENCES

[1] Bhatia, M.P.S. and Kumar, A., (2010). Paradigm shifts: from pre-web information systems to recent web-based contextual information retrieval. *Webology*, ***7***(1).

[2] Bhatia, M.P.S. and Kumar, A., (2007, November). Contextual proximity based term-weighting for improved web information retrieval. In *International Conference on Knowledge Science, Engineering and Management* (pp. 267-278). Springer, Berlin, Heidelberg.

[3] Kumar, A. and Ahmad, N., (2012). ComEx miner: Expert mining in virtual communities. *International Journal of Advanced Computer Science and Applications (IJACSA)*, ***3***(6).

[4] Kumar, A. and Bhatia, M.P.S., (2012). Community expert based recommendation for solving first rater problem. *International Journal of Computer Applications*, ***37***(10), pp.7-13.

[5] Kumar, A., &Sangwan, A. R. (2018). Expert finding in community question-answering for post recommendation. *Int. J. Eng. Technol.*, *7*(3.4), 151-159.

[6] Bogdanova, D., dos Santos, C., Barbosa, L., &Zadrozny, B. (2015, July). Detecting semantically equivalent questions in online user forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning* (pp. 123-131).

[7] Adrian Sanborn and JacekSkryzalin.(2015). Deep learning for semantic similarity.

[8] Dey, K., Shrivastava, R., & Kaushik, S. (2016, December). A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2880-2890).

[9] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., ...& Shah, R. (1993). Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(04), 669-688.

[10] Wang, G., Gill, K., Mohanlal, M., Zheng, H., & Zhao, B. Y. (2013, May). Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1341-1352).

[11] Saedi, C., Rodrigues, J., Silva, J., Branco, A., &Maraev, V. (2017, August). Learning profiles in duplicate question detection. In *2017 IEEE international conference on information reuse and integration (IRI)* (pp. 544-550). IEEE.

[12] Rodrigues, J., Saedi, C., Maraev, V., Silva, J., &Branco, A. (2017, August). Ways of asking and replying in duplicate question detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)* (pp. 262-270).

[13] Chen, Z., Zhang, H., Zhang, X., & Zhao, L. (2018). Quora question pairs. *University of Waterloo*.

[14] Zhang, Y., Lo, D., Xia, X. and Sun, J.L., (2015). Multi-factor duplicate question detection in stack overflow. *Journal of Computer Science and Technology*, ***30***(5), pp.981-997.

[15] Ahasanuzzaman, M., Asaduzzaman, M., Roy, C.K. and Schneider, K.A., (2016, May). Mining duplicate questions in stack overflow. In *Proceedings of the 13th International Conference on Mining Software Repositories* (pp. 402-412). ACM.

[16] Silva, R.F., Paixão, K. and de Almeida Maia, M., (2018, March). Duplicate question detection in stack overflow: A reproducibility study. In *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)* (pp. 572-581). IEEE.

[17] Chandu, K.R., Chinnakotla, M., Black, A.W. and Shrivastava, M., (2017, September). Webshodh: A code mixed factoid question answering system for web. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 104-111). Springer, Cham.

[18] Knight, K. and Graehl, J., (1998). Machine transliteration. *Computational linguistics*, **24**(4), pp.599-612.

[19] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[20] Felix A G, Schmidhuber J., Cummins F. (1999). "Learning to Forget: Continual Prediction with LSTM", *Ninth International Conference on Artificial Neural Networks ICANN*. (Conf. Publ. No. 470).

[21] Kumar, A. and Joshi, A., 2017, March. Ontology Driven Sentiment Analysis on Social Web for Government Intelligence. In *Proceedings of the Special Collection on eGovernment Innovations in India* (pp. 134-139). ACM.

[22] Kumar, S., Negi, A., & Singh, J. N. (2019). Semantic segmentation using deep learning for brain tumor MRI via fully convolution neural networks. In Information and Communication Technology for Intelligent Systems (pp. 11-19). Springer, Singapore.

[23] Kumar, S., Negi, A., Singh, J. N., & Verma, H. (2018, December). A deep learning for brain tumor mri images semantic segmentation using fcn. In 2018 4th International Conference on Computing Communication and Automation (ICCCA) (pp. 1-4). IEEE.

[24] Kumar, S., Singh, J. N. and Kumar, N. (2020). An Amalgam Method efficient for Finding of Cancer Gene using CSC from Micro Array Data. International Journal on Emerging Technologies, **11**(3): pp. 207–211.

[25] Imtiaz, Z., Umer, M., Ahmad, M., Ullah, S., Choi, G. S., & Mehmood, A. (2020). Duplicate questions pair detection using siamese malstm. IEEE Access, **8**, pp. 21932-21942.

[26] Zhang, W. E., Sheng, Q. Z., Lau, J. H., Abebe, E., & Ruan, W. (2018). Duplicate detection in programming question answering communities. ACM Transactions on Internet Technology (TOIT), **18**(3), pp. 1-21.

[27] Baltes, S., & Treude, C. (2020, June). Code duplication on stack overflow. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results (pp. 13-16).

[28] Abishek, K., Hariharan, B. R., & Valliyammai, C. (2019). An enhanced deep learning model for duplicate question pairs recognition. In Soft Computing in Data Analytics (pp. 769-777). Springer, Singapore.

[29] Kumar, A., & Bhatia, M. P. S. (2012). Community expert based recommendation for solving first rater problem. International Journal of Computer Applications, **37**(10), 7-13.

[30] Zhang, Y., Lo, D., Xia, X., & Sun, J. L. (2015). Multi-factor duplicate question detection in stack overflow. Journal of Computer Science and Technology, **30**(5), 981-997.

[31] Brooke, J., Tofiloski, M., & Taboada, M. (2009, September). Cross-linguistic sentiment analysis: From English to Spanish. In Proceedings of the international conference RANLP-2009 (pp. 50-54).

[32] Dey, K., Shrivastava, R., & Kaushik, S. (2016, December). A paraphrase and semantic similarity detection system for user generated short-text content on microblogs. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2880-2890).

[33] Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C.& Shah, R. (1993). Signature verification using a "siamese" time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence, **7**(04), pp. 669-688.

[34] Saedi, C., Rodrigues, J., Silva, J., Branco, A., & Maraev, V. (2017, August). Learning profiles in duplicate question detection. In 2017 IEEE international conference on information reuse and integration (IRI) (pp. 544-550). IEEE.

[35] Homma, Y., Sy, S., & Yeh, C. (2016). Detecting duplicate questions with deep learning. In Proceedings of the International Conference on Neural Information Processing Systems (NIPS).