Deep learning-based approach for detecting similar questions on stack overflow

Mrs. SK. KARIMUNNI[1], K. VENKATA PADMAVATHI[2]

[1] Assistant Professor of MCA, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

[2]PG Scholar, Dept of MCA, Audisankara College of Engineering and Technology (AUTONOMOUS) Gudur (M), Tirupati (Dt), AP

ABSTRACT_ Stack Overfow is a popular Community-based Question Answer (CQA) website focused on software programming and has attracted more and more users in recent years. However, similar questions frequently appear in Stack Over flow and they are manually marked by the users with high reputation. Automatic duplicate question detection alleviates labor and effort for users with high reputation. Although existing approaches extract textual features to automatically detect similar questions, these approaches are limited since semantic information could be lost. To tackle this problem, we explore the use of powerful deep learning techniques, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM), to detect similar questions in Stack Over_ow. In addition, we use Word2Vec to obtain the vector representations of words. They can fully capture semantic information at document-level and word-level respectively. Therefore, we construct three deep learning approaches WV-CNN, WV-RNN and WV-LSTM, which are based on Word2Vec, CNN, RNN and LSTM, to detect similar questions in Stack Over_ow. Evaluation results show that WV-CNN and WV-LSTM have made signi_cant improvements over four baseline approaches (i.e., DupPredictor, Dupe, DupPredictorRep-T, and DupeRep) and three deep learning approaches (i.e., DQ-CNN, DQ-RNN, and DQ-LSTM) in terms of recall-rate@5, recall-rate@10 and recall-rate@20. Furthermore, the experimental results indicate that our approaches WV-CNN, WV-RNN, and WV-LSTM outperform four machine learning approaches based on Support Vector Machine, Logic Regression, Random Forest and eXtreme Gradient Boosting in terms of recall-rate@5, recall-rate@10 and recall-rate@20.

1.INTRODUCTION

Quora1, Yahoo! Answers2, and Stack Overflow are just a few examples of the growing popularity of CQA websites where users can ask and answer questions in a community setting. 3 To answer common questions about computer

programming, check out Stack Overflow. On Stack Overflow, anyone can pose a question at any time. Over 18 million questions were posted to Stack Overflow as of October 2019. Even though detailed posting ethical guidelines were provided, many posed questions are of low quality [1]. Many questions asked in Stack Overflow have already been asked and answered, even though users are prompted to explore forums before submitting new questions. Users with high reputation on Stack Overflow manually mark the duplicate questions in an attempt to limit the amount of duplicate questions, which requires a significant investment of time and energy from the community at large. On top of that, it takes a considerable amount of time until a sizable number of duplicate questions are uncovered. More than 65% of duplicate questions required at least one day to close, and a major part of duplicate questions are closed after one year [2], as reported by Ahasanuzzaman et al. Therefore, a method for automatically detecting duplicate queries on Stack Overflow is needed. The process of automatically detecting duplicate questions on Stack Overflow has been investigated in prior publications. Taking into account the similarity features of themes, titles, descriptions, and tags of each question pair, Zhang et al. suggested a DupPredictor approach to automatically detect duplicate questions in Stack Overflow [3]. Duplicate questions on Stack Overflow can be easily identified with the help of a method created by Ahasanuzzaman et al., called Dupe [2]. Each of the five features—cosine similarity value, term overlap, entity overlap, entity type overlap, and wordNet similarity—are individually important. Based on DupPredictor [3] and Dupe [2], Silva et al. developed two replication methods, DupPredictorRep-T and DupeRep, to identify duplicate questions on Stack Overflow [4]. However, while these methods have successfully automated the process of identifying Stack Overflow duplicate questions, they have some limitations due to the potential loss of semantic information. Many NLP tasks, including text categorization [5] and sentiment analysis [6], are now being performed using either classic machine learning methods or more recent deep learning methods. At times, the more conventional machine learning methods outperformed their deep learning counterparts. Yet deep learning has also been used to the resolution of certain software engineering problems, such as the detection of code clones [7], the detection of bug reports [8], the prediction of semantically linkable information [9], and the prediction of software defects [10]. Some software

engineering activities have shown promising results with their use [11].

2.LITERATURE SURVEY

2.1 D. CorreaandA.Sureka,``Chafffromthewheat:Characterizationand modeling ofdeletedquestionsonstackover_ow,''in*Proc.23rdInt.Conf. WorldWideWeb(WWW)*, Jan.2014,pp.631_642.

Stack Overflow has 2.05M users, 5.1M questions, and 9.4M answers. Stack Overflow offers specific posting guidelines and an active moderator community. Despite clear communications and protections, Stack Overflow queries can be off-topic or low-quality. Experienced community members and moderators can eliminate such questions. We investigate deleted Stack Overflow queries. Our analysis has two parts: I Characterizing deleted questions during 5 years (2008-2013), and (ii) Predicting deletion at question creation. Our work characterises question deletion phenomena. Over time, more questions are eliminated. Once a question is voted to be deleted, the community acts quickly. Authors delete questions to save reputation. We occasionally accidentally delete good queries, but they're promptly re-upvoted. Deleted questions are at the bottom (lowest quality) of Stack Overflow's question pyramid. We also create a prediction algorithm to detect question deletion. We test 47 user profile, community-generated, question content, and syntactic style aspects and get 66% correctness. All four feature categories are essential for prediction, according to our analysis. Our findings offer ways to maintain content quality on Q&A websites.

2.2 Y.Zhang,D.Lo,X.Xia,andJ.-L.Sun,Multi-factorduplicatequestion detection instackover ow" 2015 Sep;30(5):981-997 J.Comput.Sci.Tech.

Software developers contribute their knowledge on Stack Overflow. Stack Overflow questions may be duplicates if they express the same idea. Duplicate questions make Stack Overflow site upkeep harder, waste resources, and make developers wait for already-available answers. Stack Overflow lets users manually mark duplicate queries. Manually spotting duplicate Stack Overflow questions is a difficult task. Detecting duplicate questions requires an automated solution. In this research, we offer DupPredictor, an automated technique that takes a new question as input and finds potential duplicates using various parameters. DupPredictor pulls a question's title, description, and tags. Title, description, and tags are required when posting a question. Using a topic model,

DupPredictor computes each question's latent subjects. Next, it compares titles, descriptions, latent subjects, and tags for each pair of questions. These four similarity scores are used to provide a comprehensive new score. To test DupPredictor, we used a Stack Overflow dataset with 2 million queries.

3.PROPOSED SYSTEM

These days, all programmers use Stack Overflow to ask questions and receive answers. Because this service is used by people all over the world, a large number of questions will accumulate, some of which will be duplicates. To remove these duplicates, highly experienced people will analyse the questions and mark them as duplicates (non-master questions), with the unique questions being considered master questions. However, this method requires a lot of human labour, so the author has modified three algorithms to detect duplicate questions from Stack Overflow.Because they lack semantic similarity, the three current methods are not good enough to obtain good prediction recall or accuracy.

The author of the proposed work is using the WORD2VEC method to alter three algorithms: CNN (convolution neural networks), LSTM (long short term memory), and RNN (recurrent neural networks). After the WORD2VEC algorithm transforms data into an integer vector with semantic similarity, the vector is fed to the three algorithms mentioned above to create a training model, which is then tested using test data to determine its prediction accuracy or recall. Among the three algorithms mentioned above, LSTM has superior recall value and question detection/prediction accuracy.

3.1 IMPLEMENTATION

1) Upload Stack Overflow Dataset: using this module we will upload questions dataset to application and pre-process question to remove special symbols and stop words.

2) Convert Dataset to Word2Vec: Using this module we will convert question dataset into integer vector representation by using PYTHON built in class called Count Vectorizers.

3) Run RNN Algorithm: Word2vec data will be passed to RNN algorithm to generate training model and then this model will be applied on test data to calculate recall and accuracy.

4) Run CNN Algorithm: Word2vec data will be passed to CNN algorithm to generate training model and then this model will be applied on test data to calculate recall and accuracy.

5) Run LSTM Algorithm: Word2vec data will be passed to LSTM algorithm to

generate training model and then this model will be applied on test data to calculate recall and accuracy.

6) Recall graph: Using this module we will show comparison graph between all algorithms.

7) Detect Duplicate Questions Test File: Using this module we will upload test questions and then apply train model on this test question to detect whether question is Master Question or Non-Master Question

4.DATASET

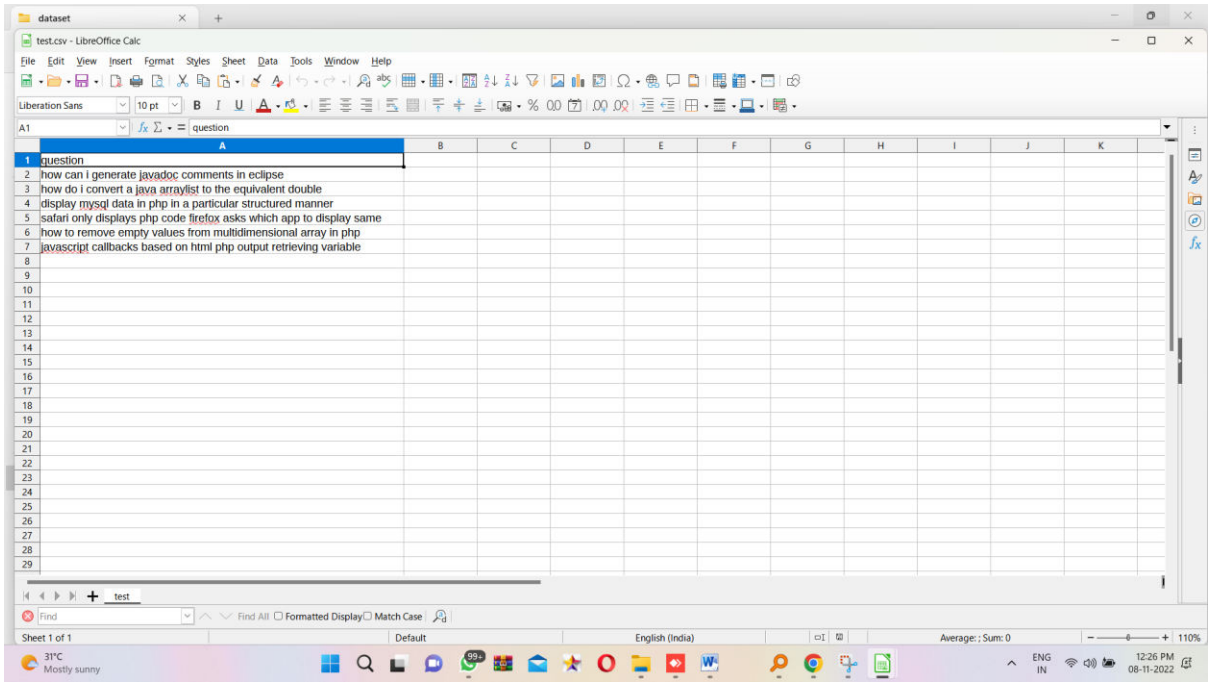

Fig 1:Dataset Values

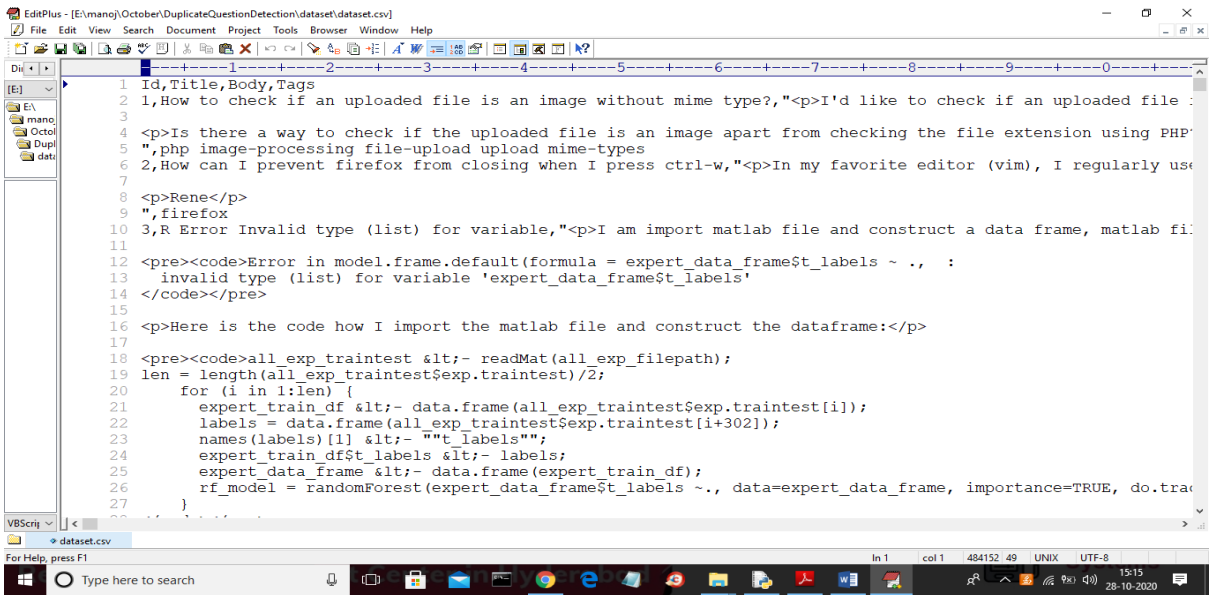

Fig 2:In above screen in dataset we have columns like Id, Title, Body and Tags and by using above dataset we will train models. In this dataset if question is duplicate then we can see tag with name as 'Possible Duplicate' and we will inform such question to model to treat as duplicate question and in new question if such duplicate question words appear then model will predict or detect as duplicate question.

5.RESULTS AND DISCUSSION

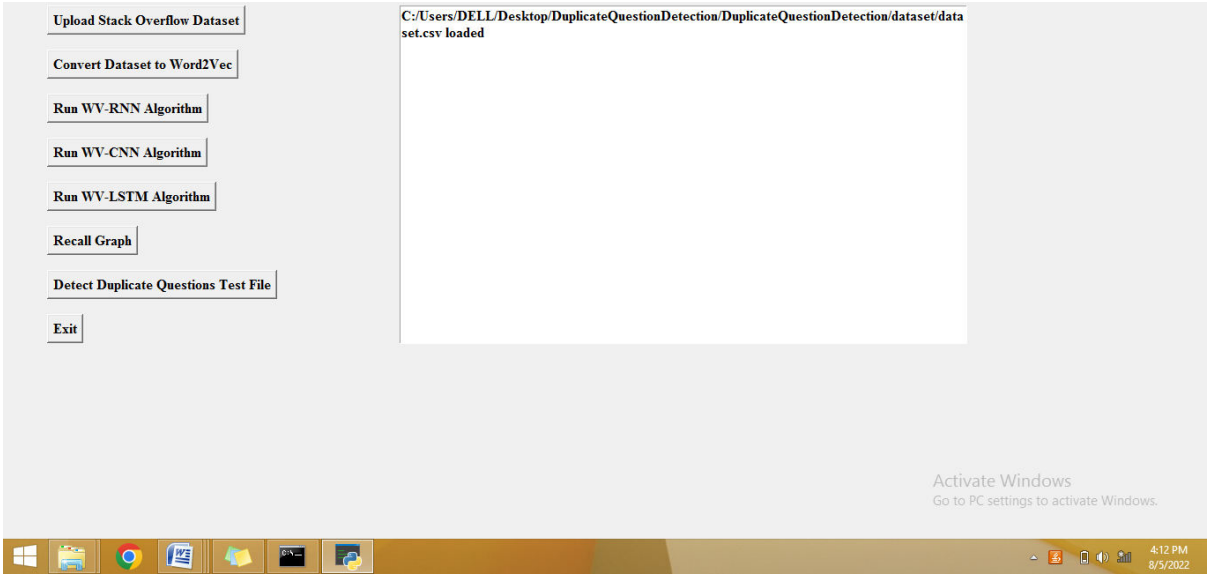

Fig 3:In above screen dataset loaded and now click on 'Convert Dataset to Word2Vec' button to convert dataset question into vector representation

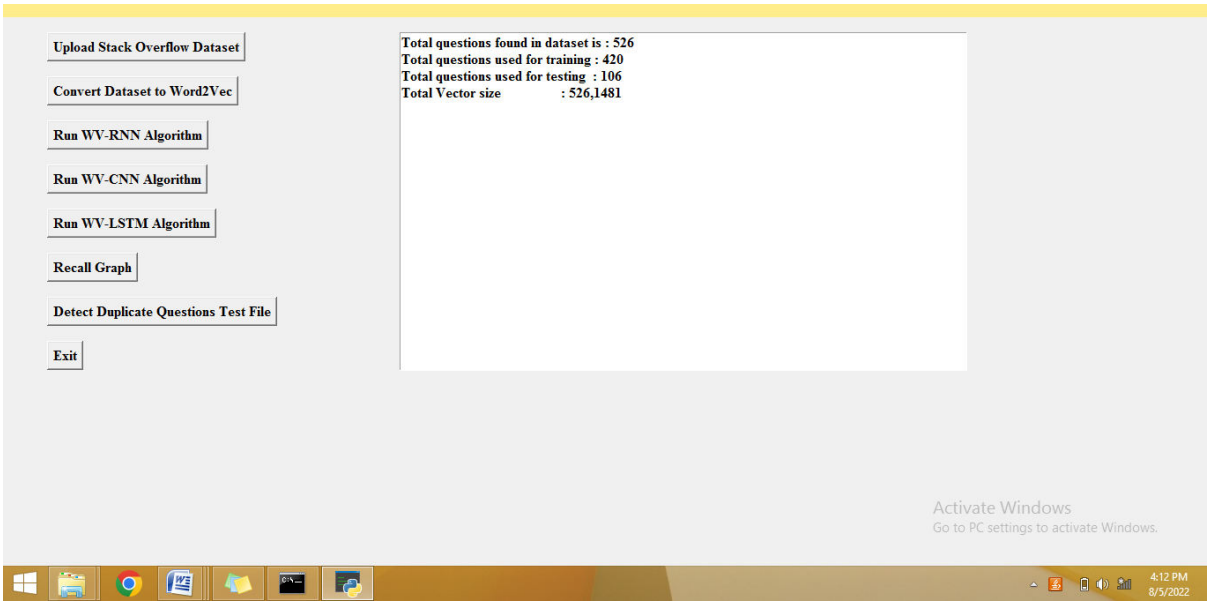

Fig 4:In above screen in dataset total 526 questions are there and application using 420 questions for training and 106 questions for testing and then generate a vector which contains 526 rows and 1481 columns where rows represents question number and column represents words count from that question. Now word2vec is ready and now click on 'Run WV-RNN Algorithm' button to train RNN on Word2Vec data

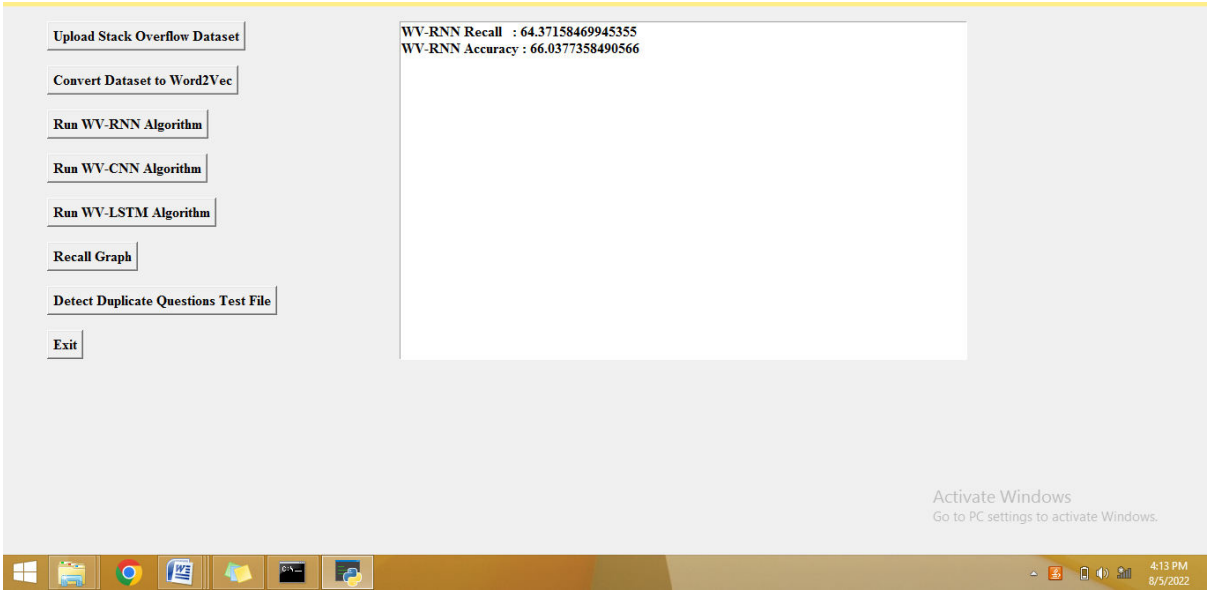

Fig 5:In above screen after applying Word2Vec on RNN we got prediction/detection accuracy as 62% and recall as 61% and now click on 'Run WV-CNN Algorithm' to get its accuracy value

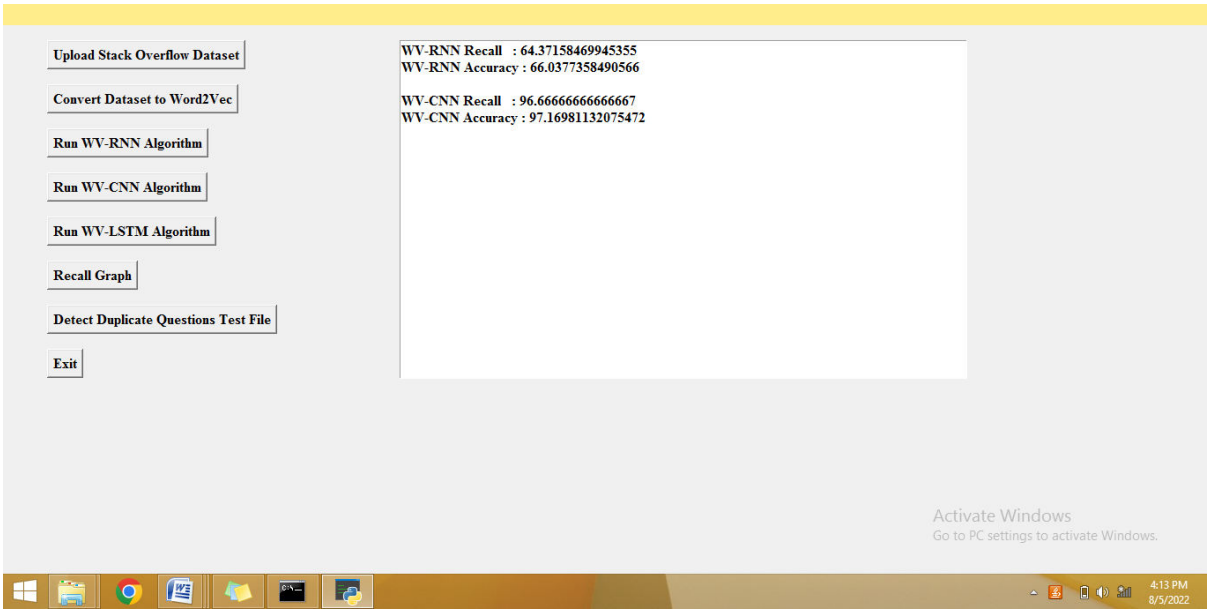

Fig 6:In above screen WV-CNN got more than 90% accuracy and recall and now click on 'Run WV-LSTM Algorithm' button to get its accuracy

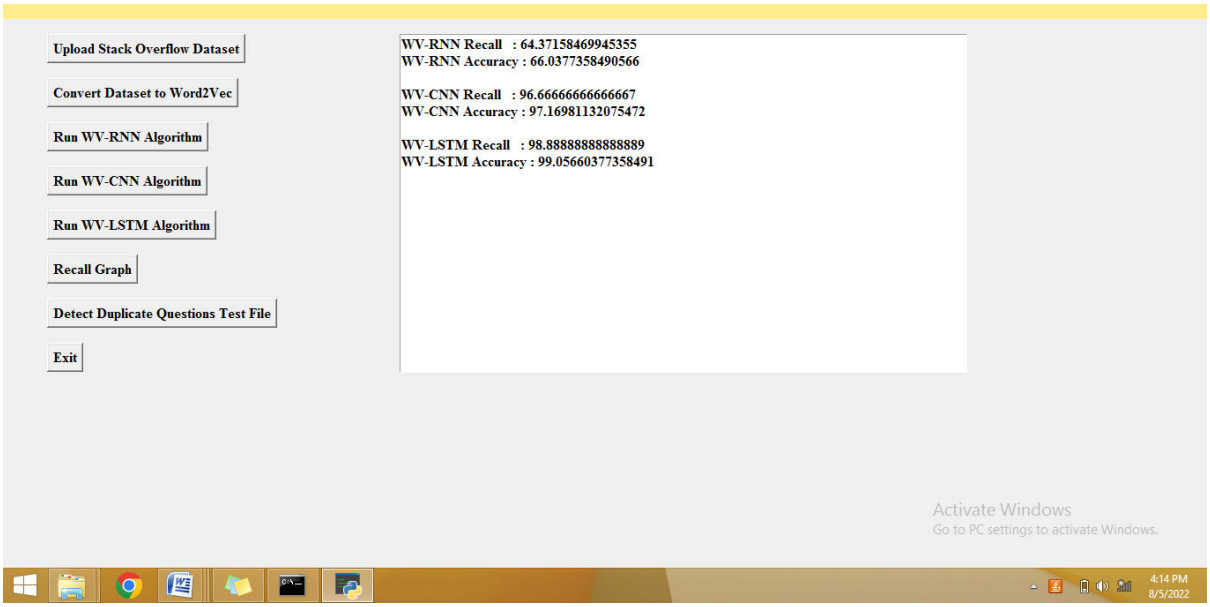

Fig 7:In above screen LSTM got nearly 99% accuracy and recall values and in black console we can see MODEL details

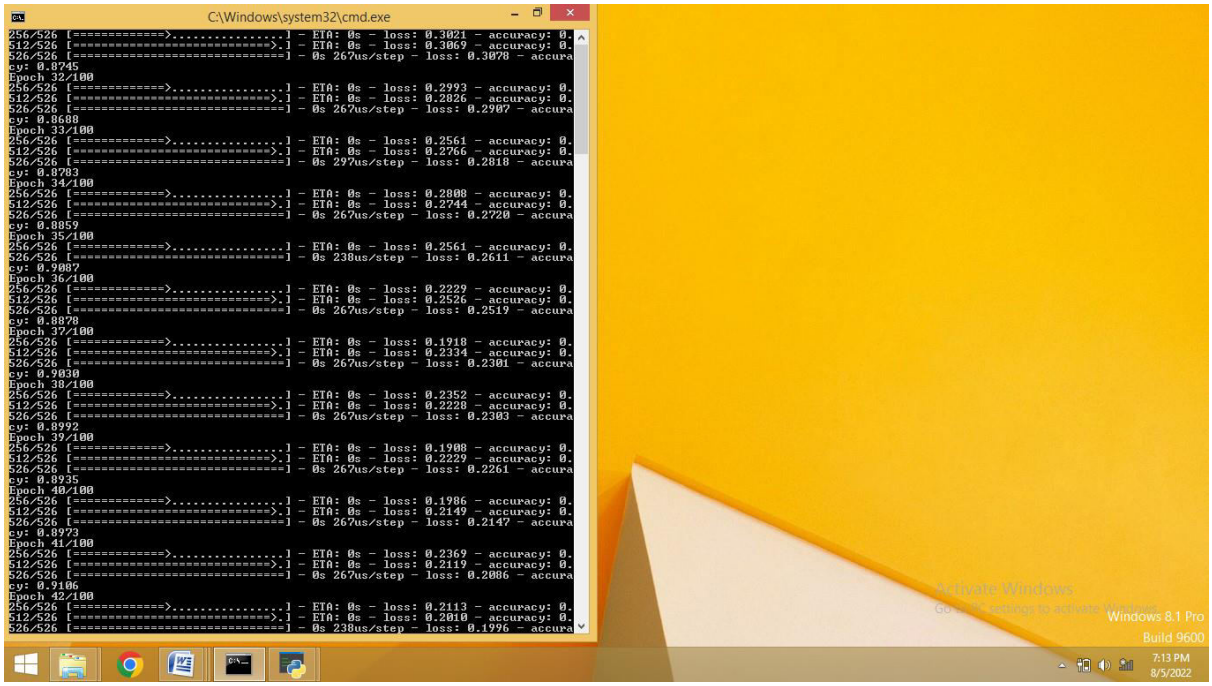

Fig 8:In above screen we can see LSTM model details and now click on 'Recall Graph' button to get below graph

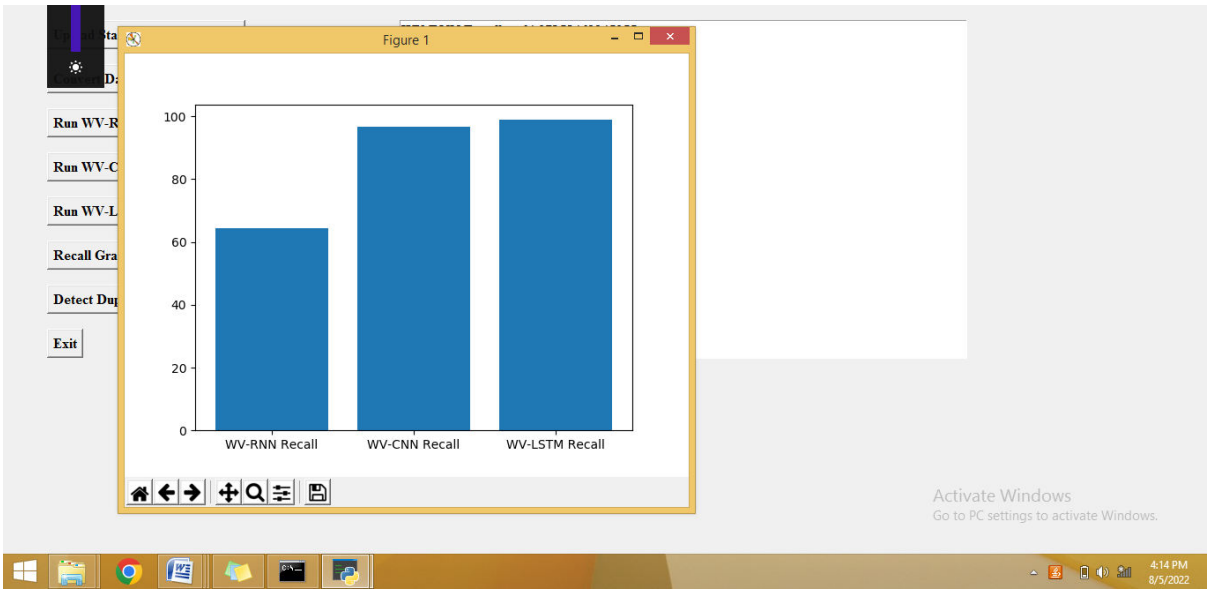

Fig 9:In above graph x-axis represents algorithm name and y-axis represents recall value and from above graph we can conclude that LSTM is performing well. Now click on 'Detect Duplicate Questions Test File' button to upload test file and then model will detect whether question is master or non-master question

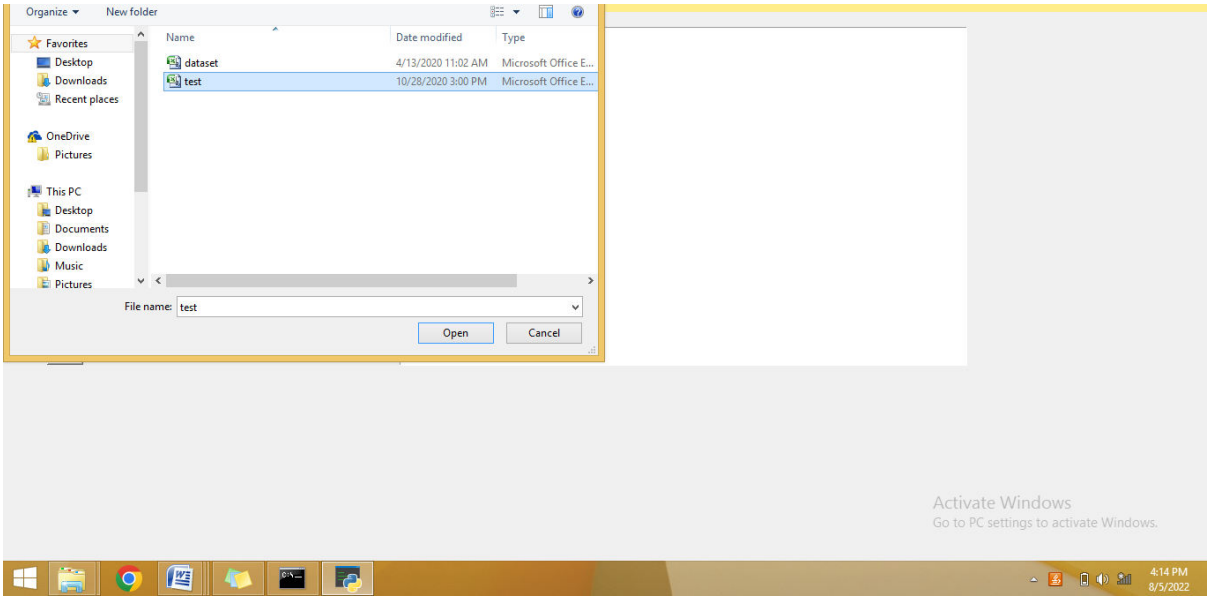

Fig 10:In above screen uploading test.csv file and then click on 'Open' button to detect whether test.csv questions or master or non-master. Here Non-master mean duplicate and master means unique

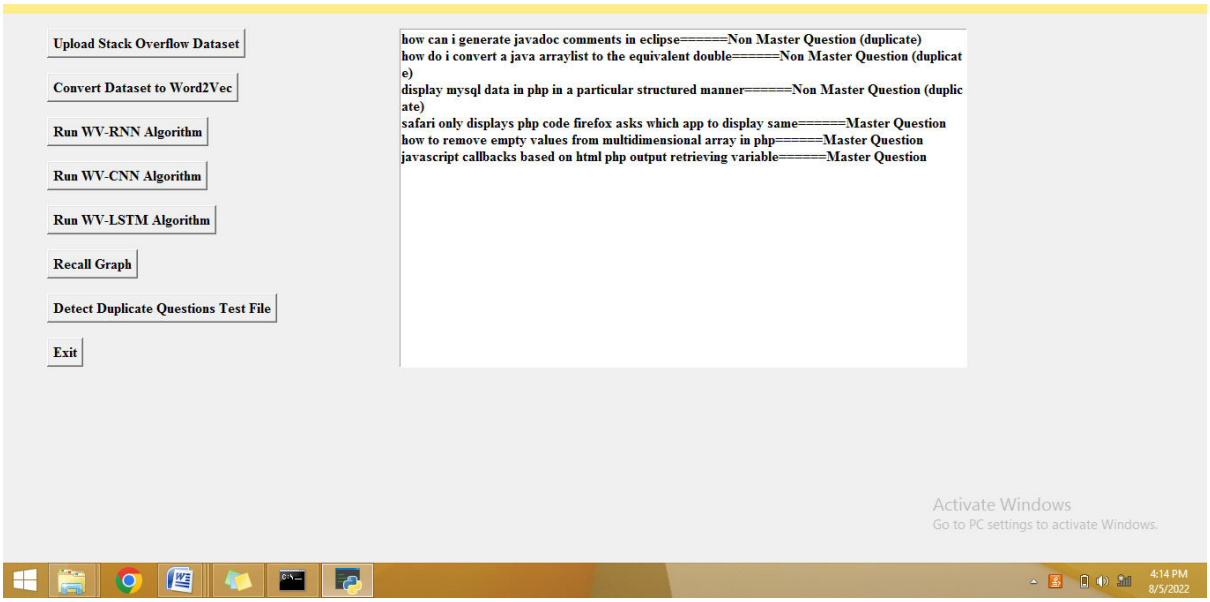

Fig 11:In above screen before equals to symbols is the question and after equals to symbol is the detection as master or non-master. This propose paper algorithms obtained recall and accuracy 99%

6.CONCLUSION

using deep learning and Word2Vec, we can identify duplicate queries on Stack Overflow. The issue of duplicate question detection on Stack Over ow is examined using three distinct deep learning

approaches: convolutional neural networks, recurrent neural networks, and long short-term memory. Word2Vec is also employed for the extraction of word vectors. Based on Word2Vec, CNN, RNN, and LSTM, this study develops three deep learning methods, WV-CNN, WV-RNN, and WV-LSTM, to identify duplicate questions in Stack Over ow. Each question combination on Stack Over ow can have its whole semantic information captured, from the individual words to the entire context of the document. Compared to four baseline methods (i.e. DupPredictor, Dupe, DupPredictorRep-T, and DupeRep) and four machine learning methods (i.e. SVM, LR, RF, and Xgboost), the recall-rate@5, recall-rate@10, and recall-rate@20 are all significantly higher for our approaches WV-CNN and WV-LSTM across all six question sets. Further, our approaches WV-CNN, WV-RNN, and WV-LSTM show significant gains over three deep learning approaches (i.e., DQ-CNN, DQ-RNN, and DQ-LSTM) across six distinct question groups in terms of recall- rate@5, recall-rate@10, and recall-rate@20

REFERENCES

[1] BuljanM,Bulana V and Sandra S 2008Variation inClinicalPresentation ofBasalCellCarcinoma (Kroasia: University Department of Dermatology and Venereology Zagreb Croatia) p 25-30.

[2] Cipto H, Suriadiredja AS. Tumor kulit. Dalam: Menaldi SL, Bramono K, Indriatmi W, editor. Ilmu penyakit kulit dan kelamin. Edisi ketujuh. Jakarta: Badan Penerbit FKUI; 2016. h.262-276.

[3] TeresiaR, Savera,Winsya H, Suryawan and AgungWS 2020Deteksi Dini KankerKulit Menggunakan K-NN dan Convolutional Neural Network J. JTIIK. 7 2 p 373-378.

[4] Md Ashraful Alam Milton 2018 Automated Skin Lesion Classification Using Ensemble of DeepNeural Networks in ISIC: Skin Lesion Analysis Towards Melanoma Detection Challenge.

[5] Serban Radu SJ, Loretta Ichim, et al 2019 Automatic Diagnosis of Skin Cancer Using Neural Networks (Bucharest, Romania: The XIth International Symposium on Advanced Topics in Electrical Engineering March 28-30).

[6] Nazia Hameed, et al Multi-Class Skin Diseases Classification Using Deep Convolutional Neural Networkand Support Vector Machine.

[7] Xinyuan Zhang, et al 2018 TowardsImproving Diagnosis of Skin Diseases by Combining Deep Neural Network and Human Knowledge BMC Medical Informatics and Decision Making 18(Suppl 2) p 59

[8] Khalid M. Hosny, et al 2019 Classification of skin lesions using transfer learning and augmentation with Alex-net PLOS ONE.

[9] Marwan AA 2019 Skin Lesion Classification Using Convolutional Neural Network With Novel Regularizer IEEE Access.

[10] The International Skin Imaging Collaboration (ISIC).Accessed: June 2020. [Online]. Available: https: //www.isicarchive.com/#!/topWithHeader/ onlyHeaderTop/gallery [11] Mousumi Roy, et al 2016 Dermatofibroma: Atypical Presentations Indian J. Dermatology.

[12] R. Delila Tsaniyah, Aspitriani and Fatmawati "Prevalensi dan Gambaran Histopatologi Nevus Pigmentosus di Bagian Patologi Anatomi Rumah Sakit Dr. Mohammad Hoesin Palembang," Periode 1 Januari 2009-31 Desember 2013.

[13] Syril Keena T. Cutaneous squamous cell carcinoma. Journal of The American Academy of Dermatology. Volume 78, Issue 2, p237-432, e33-e55

[14] Marco Rastrelli, et al 2014 Melanoma: Epidemiology, Risk Factors, Pathogenesis, Diagnosis, and Classification 28 no. 6 p 1005-1011

[15] P.Kim, 2017 MATLAB Deep Learning: With Machine Learning, Neural Networks, and Artificial Intelligence.

[16] A. F. Agarap 2008 Deep Learning using Rectified Linear Units (ReLU) 1 p 2–8 [Online]. Available: http://arxiv.org/abs/1803.08375.

[17] S. Khan, H. Rahmani, S. A. A. Shah, M. Bennamoun, G. Medioni and S. Dickinson 2018 A Guide to Convolutional Neural Networks forComputer Vision. Morgan Claypool [Online] Available: https: //ieeexplore.ieee.org/document/8295029

[18] H. Robbins and S. Monro 1985 A Stochastic Approximation Method" in Springer [19] J. Duchi, E. Hasan and Y. Singer 2011 Subgradient Methods for Online Learning and Stochastic Optimization J. of Machine Learning Research

[20] Y. Yunlong, L. Fuxian. 2019 Effective Neural Network Training with a New Weighting Mechanism-Based Optimition Algorithm IEEE Access.

Author Profiles

Mrs. SK. KARIMUNNI currently she has working Assistant Professor in Audisankara College of Engineering & Technology Gudur(M), Tirupati (DT), She is done M.Tech from Quba College of Engineering and Technology, Venkatachalam in 2015.

K. VENKATA PADMAVATHI is pursuing MCA from Audisankara college of Engineering &Technology (AUTONOMOUS), Gudur, Affiliated to JNTUA in 2024. Andhra Pradesh, India.