2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-6654-5262-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCNT54827.2022.9984492

Quora Question Pairs Identification and Insincere Questions Classification

Sai Surya Teja Gontumukkala, Yogeshwara Sai Varun Godavarthi, Bhanu Rama Ravi Teja Gonugunta, Deepa Gupta* ,Suja Palaniswamy*
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
saisurya029@gmail.com, gysaivarun31@gmail.com, ravitejagonugunta150@gmail.com,g_deepa@blr.amrita.edu*,p_suja@blr.amrita.edu*

***Abstract* — Quora is a question-answering site where people ask questions and reply to the existing questions which makes Quora a great interactive platform but it also has few challenges such as the occurrence of duplicate questions which lead to ambiguity and insincere questions that degrade the value of the site. In this research work, we have proposed a method to overcome these two challenges by using techniques of Natural Language Processing (NLP) and Deep Learning (DL). Five different word embeddings were used for both the problems, Bi-directional Long Short-Term Memory (BiLSTM) and Bi-Gated Recurrent Unit (BiGRU) architecture with attention mechanism were used for insincere question classification and Siamese Manhattan Long Short-Term Memory (MaLSTM) architecture were used for question pairs identification. The implemented models are performing well, in terms of accuracy, precision, recall, and F1 Score. Our research work has achieved the highest accuracy of 90% and highest F1 score of 0.89 by using Paraphrase-MiniLM-L6-v2 + Siamese MaLSTM for Quora Question Pairs Identification and for Insincere Questions Classification our model achieved the highest accuracy of 95% and highest F1 score of 0.82 by using FastText + BiLSTM + BiGRU. Our results were compared with literature and our research work has outperformed baseline models.**

Keywords—NLP, Word2Vec, GloVe, FastText, Doc2Vec, Paraphrase-MiniLM-L6-v 2, SiameseMaLSTM, BiLSTM, BiGRU, Quora, Question, Insincere.

I. INTRODUCTION

Quora is a website where one can study and share knowledge by asking questions to others and also answering the questions which were posted by other people which entirely makes Quora, a multiple unique perspectives sharing platform[1]. It is one of the most popular websites for posing and answering questions. The users here generate, update, and organize the questions and answers. One can ask questions under a variety of categories [2]. Understanding the meaning of sentences and how they are related to each other will make understanding the content generated by a user on Quora. Due to the large number users who are using Quora, somewhere or the other perspective of two users might match i.e., they might ask the same or related questions for the same purpose, this needs to duplicate questions on a website [3]. Identifying duplicate questions or question pairs simplifies the process to find better replies or answers for the question and result are saving the time which in turn makes the user experience better for writers and users. This helps the user identify which category the query belongs to [4]. Another problem with this kind of forum is insincerity i.e., some people take advantage of this website in the wrong manner. They ask questions that should not be posted on an international website, which is one of the problems that could affect the quality of the internet forums. This problem spreads hatred among people about an individual or group of individuals[5-8].

As a result, our work aims to address the problem of duplicate questions or question pairs identification using NLP and DL. Addressing these two problems would be very helpful in helping internet forums to organize their database and improve the value of the question forums.

Semantic matching is a process in NLP that finds semantically related information i.e., identifying if two sentences have the same meaning [9]. Word Embeddings are real-valued vectors that represent the meaning of a word. It is inferred that words with equivalent vector spaces will have similar meanings [10]. In machine learning, classification is a technique used to categorize new observations based on training data. Those classification jobs with only two class labels are referred to as binary classification [11]. Quora question pairs identification is a semantic matching binary classification problem, where word embeddings can be used to find out whether two questions are similar or not. Quora insincere question classification is a binary classification problem, where word embeddings will be used to train the model to predict whether the question is sincere or not.

Long Short-Term Memory (LSTM)s are specifically invented to address the long-term dependency issue. Bidirectional long short-term memory (bi-LSTM) is a variety of LSTM models that processes data in both directions. The BI-LSTM is distinguished from other LSTMs by its ability to flow data in both directions. The attention mechanism helps in looking back at all the hidden states for making the prediction i.e., it helps in searching the most relevant information.

In this paper, BiLSTM + BiGRU with an attention mechanism is used to predict whether the given question is insincere or not. This combination of BiLSTM and BiGRU gives the model the power to handle more independent and long-range characteristics of the sentences. Siamese networks are networks that have two or more identical sub-networks in them. MaLSTM is a type of LSTM where Ma stands for Manhattan distance and Siamese MaLSTM is used to predict whether the questions are pair or not, where Manhattan distance between two sentences is calculated to know whether the sentences are similar or not which is a very efficient way to determine the context similarity between two sentences.

Our major contributions are as follows:

- Investigated five different embeddings namely Word2Vec, GloVe, FastText, Doc2Vec, and Paraphrase-MiniLM-L6-v2 for Quora question pairs identification and insincere questions classification.

- Developed a Siamese MaLSTM model for question pairs identification and a BiLSTM + BiGRU model with an attention mechanism for insincere question classification by using five different embeddings.
- Compared all the models and identified the best one out of them.

The paper is organized as follows: Related works were briefly and precisely discussed in Section - II. The details related to the datasets were discussed in Section – III. The methodology, proposed system architecture along with concepts of word embeddings, BiLSTM, MaLSTM, and BiGRU are explained briefly in Section - IV. A walk-through of experimented Rresults was given in Section - V. In Section - VI, Conclusion, and Future Scope are discussed.

II. Related work

To identify the quora question pairs two approaches were used, Siamese architecture with the single LSTM and with two LSTM. The GloVe is used for Word Embeddings. One of the models is a Siamese network i.e., a single LSTM is being used for both sentences and the other one is a two LSTM network i.e., it sends the two sentences in sequence [1]. To identify the duplicate questions, they proposed a Siamese MaLSTM network and worked with 3 different embeddings such as Word2Vec, FastText, and FastText-Sub word, along with these three they have used a blend of these three embeddings that outperformed all the others and gave benchmark results [2]. The stage of the BERT where it is pre-trained by joining left and right contexts gave itself the success it has today in understanding language very intimately. This understanding is very useful in this classification task. The authors in [3] made a comparative study between RNN+GRU, RNN+LSTM+GRU, and BERT. Out of these, BERT gave a good F1 score of 0.70 for insincere questions classification.

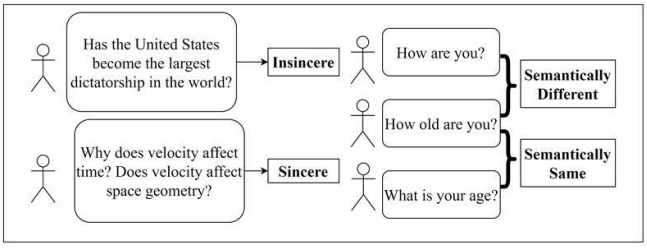
The data was precisely understood and explored; machine learning algorithms were used to solve the Quora question pairs identification problem. Bag of words neural network gave the best output [4]. A classification approach that is based on deep neural network architecture with BiLSTM, ReLu, and Sigmoid Activation function which gave 95.25 percent accuracy [8]. By combining several techniques such as BERT, rules made by humans, and labels being given by humans to improve the accuracy of true labels of the Quora question pairs identification dataset [12]. A deep learning approach was proposed which classifies insincere questions by using word embeddings [13]. A variety of machine learning and neural network algorithms were used to classify the insincere questions and the results were analyzed [14]. Multiple supervised ML algorithms were used to classify the insincere questions and the Decision Tree was chosen to be the best out of them [15].

Examination of the task of identifying insincere questions in Quora data was done using machine learning [16]. Six different machine learning models were used to classify the insincere questions and it was found that the multi-layer perceptron achieved the highest accuracy score of 87.8 out of all of them [17]. Different Machine Learning Models were used to identify the duplicate questions in Quora data in which Siamese LSTM outperformed all the other models by giving a very less loss of 0.21[18]. Machine learning algorithms like Logistic Regression, SVM, and Gradient boosted decision trees along with hand-picked features and word embeddings were used to predict whether the questions were pairs or not, out of which decision trees gave less log loss compared to the other models [19].

Keeping BiLSTM as a baseline, four different deep learning models were proposed out of which the model which consisted of framed GBDT gave the best accuracy and proved that it can handle lexical gaps [20]. Multi-head attention process was used to detect the duplicate questions which gave an accuracy of 89 % [21]. Different ML algorithms were used to identify the question pairs, XG boost algorithm gave the best accuracy of 83% out of all [22]. A support vector classifier with the help of precomputed features is used to calculate the similarity between the two sentences to find whether they are a pair or not [23].

The previously done work has either worked on question pairs identification or insincere question classification, tried with only one embedding technique and used only word embeddings. In this research work, both the issues i.e., question pairs identification and insincere question classification have been addressed, five different embeddings are used for solving both the problems and also tried to improve the performance of the model by using State of the Art Models like BiLSTM + BiGRU which is very effective in classification task even for longer data and Siamese MaLSTM which is proven to the most efficient way to identify the semantic similarity between two sentences. In addition to this, our work used sentence embeddings (Doc2Vec) and sentence transformer (Paraphrase-MiniLM-L6-v2) along with the word embeddings. Sentence embeddings and sentence transformers are proven to be powerful and provide more accurate results when the data consists of long sentences.

III. Data Source



(a) (b)

Fig.1. Example for Insincere Question (a) and Questions Pairs (b)

TABLE I. DATASET DESCIRPTION

	Quora Question Pairs	**Quora Insincere Questions classification**
Number of sentences in the training data	404291	1048576
Number of sentences in the test data	1048576	375807

Both datasets i.e., Quora Question Pairs Identification[1] and Quora Insincere Question Classification[2] are collected from Kaggle. The examples for insincere questions and questions pairs are shown in Fig.1. (a) and (b) respectively. The dataset Description i.e., the number of questions in the training set and testing set for both Quora Question Pairs Identification and Quora Insincere Question Classification is given in Table I.

IV. Proposed Methodology

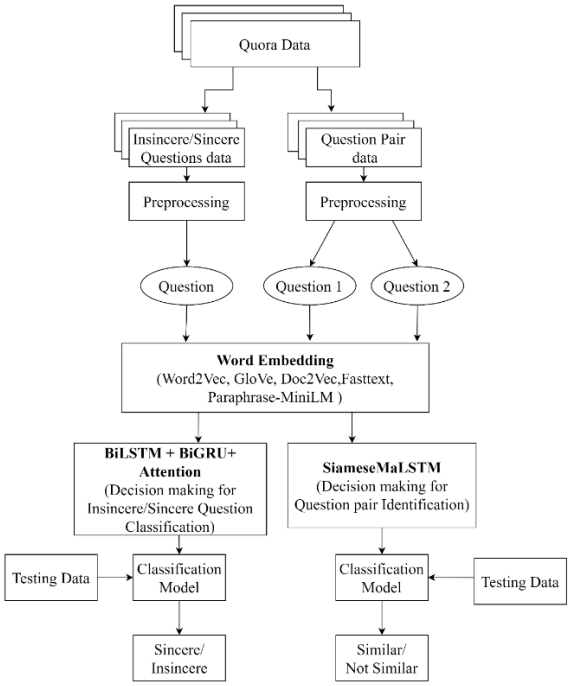


Fig. 2. Proposed System Architecture

As shown in Fig.2., The Proposed System Architecture starts with collecting the data and then applying preprocessing techniques like stop word removal, punctuation removal, misspelled word corrections, Lemmatization, etc. After that convert the sentence into vectors by using Word Embeddings. We have proposed ten frameworks for Question pairs identification and Insincere questions class. Word Embeddings used here are Word2Vec, Glove, Doc2Vec, Paraphrase, and FastText.

For decision making, Used BiLSTM + BiGRU with attention mechanism is used for Quora Question Pairs identification and Siamese MaLSTM is used for Insincere questions classification.

A. Preprocessing:

The text in the Quora questions often tends to have spelling mistakes, and punctuation because the questions are asked by common people on the website, so this text requires undergoing some preprocessing before training and converting it into vectors. Basic preprocessing techniques like stop words removal, punctuation removal, correcting the misspelled words, and lemmatization was done on the Quora data.

B. Types of Embeddings:

- Glove: GloVe expands as Global Vectors for word representation. It is a type of unsupervised learning approach developed by Stanford University which was developed to build word embeddings by combining the global word co-occurrence matrices of a particular corpus. The basic idea of developing GloVe word embedding is to make use of statistical inference to derive the relation between words. The co-occurrence matrix is different from an occurrence matrix, it gives information about how commonly a specific pair of words will co-occur with each other. A word pair that occurs with each other will be represented by the value present in the co-occurrence matrix [24].

- Word2Vec: The Word2Vec tool accepts input in a text corpus format and gives word vectors as output. It learns vector representations of words after first constructing a vocabulary from training the text data. Many NLP and machine learning systems can use the generated word vector file as a feature. Finding the nearest words for a user-specified term is a straightforward technique to investigate the learned representations [24].

- Doc2Vec: Doc2Vec embeddings by Wieting et al. are one of the best embeddings that can be used for pretraining or transfer learning model that uses semantic representations of words and wants them to be composable [25].

- FastText: FastText is a framework which is developed by one of the research labs of Facebook for learning how to build word embeddings and it can also be used for text classification. The model enables the development of an unsupervised learning algorithm or supervised learning algorithm for building the vector form representations of the words that are present in the corpus. Facebook used around 294 languages pre-trained models. FastText is extremely fast when it comes to training word vector models. In less than ten minutes, you can learn approximately one billion words. Deep neural network models can be time-consuming to train and test. To train the model, these methods employ a linear classifier [24].

- paraphrase-MiniLM-L6-v2: The paraphrase-MiniLM-L6-v2 is a sentence transformer from the hugging face library. The model adds a memory vector to a basic encoder-decoder model to capture the paragraph's topic or meaning of the input. Here the training is comparable to that of a continuous bag of words i.e., predicting a single phrase from its context. The context words in this instance are the previous words, not the adjacent ones, as they are in the form of a paragraph [25].

C. Deep Learning Models:

- BiLSTM: A BiLSTM, is a sequential data processing model which is composed of two LSTMs, one forward and one backward. BiLSTM effectively increases the quantity of data available to the network, which improves the context which will be present in the algorithm. Using bidirectional with attention mechanism will run inputs in two possible directions, one will be running from past to future and the other will be running from future to past. A BiLSTM helps in a situation like take an example "Orange is a sour fruit that is round in

shape and is orange colored.", Here the orange can be a color or it can be a fruit, just by looking at the "orange" word we cannot judge what is that, it all depends on the context of the sentence. In NLP we sometimes need to refer not only to the preceding word but also to the next word to describe a term. We use forward propagation twice in this case, once for forwarding cells and once for backward cells.

- BiGRU*:* It is also a type of RNN which consist of two gates those are reset and update gate, it works the same as an LSTM but with an output gate it is preferred for small data sets as it is computationally less costly than LSTM. A BiGRU, is a sequential data processing model which is composed of two GRU, one forward and one backward. BiGRU effectively increases the quantity of data available to the network, which improves the context which will be present in the algorithm. Using bidirectional with attention mechanism will run your inputs in two possible directions, one will be running from past to future and the other will be running from future to past.

- MaLSTM: The MaLSTM is a simple model that uses two LSTMs to determine how identical two sequences are. The variable length sequence is converted into a fixed dimensional vector embedding by the two LSTMs. The similarity measure is then computed by applying a similarity function to these vectors. The last hidden state is used as the vector embedding for the sequence. The similarity metric uses the L1 norm or the Manhattan distance. Each question is processed by the weight allocated to it. For the ultimate prediction, MaLSTM estimates the Manhattan distance. Manhattan distance outperforms alternative substitutes, such as cosine similarity, by a small margin. The reason behind selecting Manhattan distance among all the other types of similarity measures is because here the model is working on an embedding with multiple dimensions as well as a larger size.

  Manhattan distance similarity measure has been found to work well on data with high dimensions but also it has been found to take less time to compute because it calculates the absolute distance from one point to other points that lie at right angles to identify the similarity between textual attributes. The Manhattan equation is:

$$Ma = |x1 - x2| + |y1 - y2| \quad (1)$$

  In Eq. (1), x_1 and y_1 are referring to the output of the first model, x_2 and y_2 are referring to the output of the second model. The absolute difference between them refers to the similarity between the two points i.e., sentences in this case in the given input to the model.

V. Experimental Results and Analysis

All models were implemented in google colab with python 3.7 version. Siamese MaLSTM was implemented using the TensorFlow module, for all the embeddings training was done for 30 epochs and BiLSTM + BiGRU was implemented using the PyTorch module, for all the embeddings training was done for 5 epochs. Hyperparameter tuning was done for both the models by trying with the different numbers of epochs and different optimizers. It was observed that Adam optimizer outperformed the others.

The evaluation metrics used to evaluate the models are Accuracy, Precision, Recall, and F1 Score. Accuracy reveals how often the model is overall correct. The model's precision measures how well it can predict a particular category. Recall tells how frequently the model was able to identify a particular category. By calculating the harmonic mean of a classifier's precision and recall, the F1-score integrates both into a single metric. It is mainly used to compare the effectiveness of two classifiers. In the comparison of results, A refers to Accuracy, P refers to Precision, R refers to Recall and F1 refers to F1 score.

A. Parameter Setting:

TABLE II. HYPERPARAMETER TUNING OF QUESTION PAIRS IDENTIFICATION.

Parameters	Values	
	Word2Vec, GloVe, FastText, Doc2Vec	***Paraphrase-MiniLM-L6-v2***
Dataset Split	70,30	70,30
Embedding Dimension	300	385
Batch Size	64	64
Epochs	30	30
Optimizer	Adam	Adam
Loss Function	Mean Squared Error	Mean Squared Error
Number of Nodes in Dense Layer	50	50
Total number of Parameters	27, 246, 900	310, 581, 720
Number of Trainable Parameters	70, 200	87,000

TABLE III. HYPERPARAMETER TUNING OF INSINCERE QUESTIONS CLASSIFICATION.

Parameters	Values	
	Word2Vec, GloVe, FastText, Doc2Vec	***Paraphrase-MiniLM-L6-v2***
Dataset Split	70,30	70,30
Embedding Dimension	300	385
Batch Size	1000	1000
Epochs	5	5
Activation Function	ReLu	ReLu
Drop out	0.1	0.1
Optimizer	Adam	Adam
Loss Function	Mean Squared Error	Mean Squared Error
Number of Nodes in Dense Layer	16	16

Different parameters were tuned during the model training such as batch size, number of epochs, optimizer, loss function, etc. The tuned hyperparameters for question pairs identification are shown in Table II and the tuned hyperparameters for insincere questions classification are shown in Table III.

B. Result Evaluation and Comparison

The results of all the model combinations with five embeddings for question pairs identification were shown in Fig. 3. Here it is visible that the Paraphrase-MiniLM-L6-v2 sentence transformer has outperformed all the other embeddings Accuracy of 0.9, Precision of 0.85, Recall of 0.94, and F1 Score of 0.89 for question pairs identification.

The results of all the model combinations with five embeddings for insincere question classification were shown in Fig. 4. Here both Word2Vec and FastText have given the highest accuracy of 0.96, Recall of 0.83, and f1 score of 0.81 but FastText gave a precision of 0.81 which is higher than that of Word2Vec.

In Table – IV, we have compared our work with six previously done works and it can be interpreted that although there is less improvement in terms of accuracy and precision, a 7 points improvement in F1 score and 13 points improvement in Recall is achieved when compared to [15]. This signifies that our model is capable of identifying the maximum number of question pairs that are labelled as a pair correctly and it gives results that are output sensitive.

In Table – V, we have compared our work with six previously done works and it can be interpreted that although there is less improvement in terms of accuracy, 13 points of improvement in precision was achieved when compared to [8], 15 points of improvement in F1 score was achieved when compared to [7] which shows us our model is unbiased towards a label and it has been proved that it works well for imbalanced data since our data contains more questions which are labelled as sincere. 21 points improvement in Recall was achieved when compared to [11] which shows that our model is capable of classifying all the questions correctly and is output sensitive.

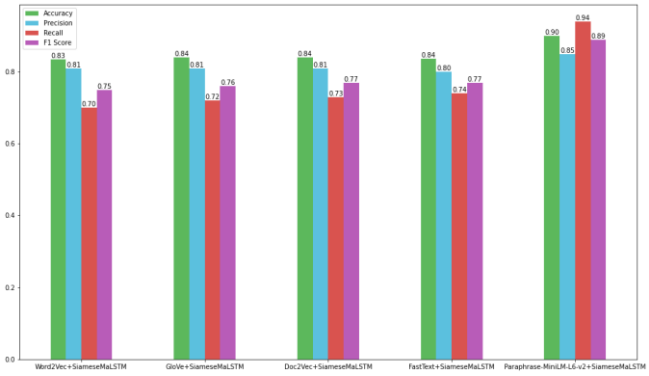


Fig. 3. Results of Question Pairs Identification

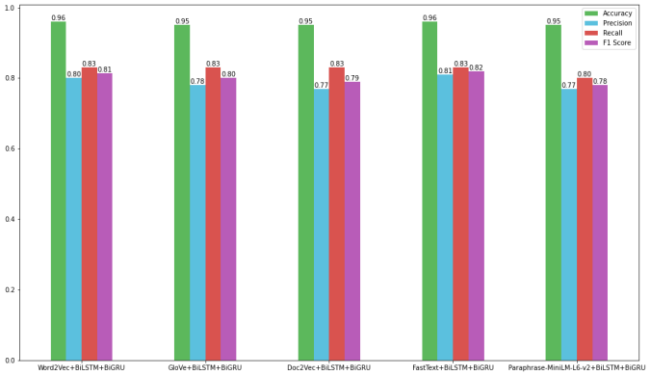


Fig. 4. Results of Insincere Question Classification

TABLE IV . COMPARISON OF RESULTS OF QUESTION PAIRS IDENTIFICATION WITH PREVIOUS WORK

References	Model	A	P	R	F1
[Dadashov et.al; 2012 [1]	LSTM	83.8%	-	-	0.79
[Zhang et.al:2018] [14]	BiLSTM + Frame-GBDT	87.92%	-	-	-
[H. Zhang, L. Chen; 2019] [15]	Neural Networks + Multi-head Attention	86.83%	0.84	0.81	0.82
Imtiaz et.al; 2020] [2]	Siamese LSTM	82.77%	0.79	0.70	0.75
[Chunamari et.al; 2020] [16]	XG Boost	81%	-	-	-
[H. T. Le et.al; 2021] [6]	BERT Model	80%	-	-	-
Proposed Model	Paraphrase-MiniLM-L6-v2 + Siamese MaLSTM	**90%**	**0.85**	**0.94**	**0.89**

TABLE V. COMPARISON OF RESULTS OF INSINCERE QUESTION CLASSIFICATION WITH PREVIOUS WORK.

Reference	Model	A	P	R	F1
[Akshay et.al; 2019] [9]	Naïve Bayes + Logistic Regression	95%	0.64	0.62	0.63
[D.Y.Kim et.al; 2019] [7]	BiLSTM	-	-	-	0.67
[Hendri, Adriani; 2019] [11]	Multilayer Perceptron	87.81%	-	-	-
[Hao Mao et.al; 2019] [8]	Averaged Perceptron	95.7%	0.68	0.59	0.63
[C.Chen; 2021] [5]	BiLSTM	95.25%	-	-	
[R. Kumar et.al; 2021] [10]	SVM	89.9%	-	-	-
Proposed Model	FastText + BiLSTM + BiGRU	**96%**	**0.81**	**0.83**	**0.82**

VI. CONCLUSION AND FUTURE SCOPE

Quora question pairs identification using MaLSTM and insincere questions classification has been done using BiLSTM + BiGRU respectively. Different embeddings were tried, out of which Paraphrase-MiniLM-L6-v2 performed the best for Question Pairs Identification and FastText performed the best for insincere question classification. We have achieved a benchmark accuracy for insincere questions classification and question pairs classification using single-word embedding.

Since the Quora insincere questions dataset has a large imbalance in the class distribution, there is a possibility of the output being sensitive towards a class. This work can be further extended by using a blend of word embeddings for question pairs identification and by removing the imbalance in the dataset for insincere question classification.

REFERENCES

[1] Dadashov, Elkhan, Sukolsak Sakshuwong and Katherin Yu, Quora Question Duplication, 2017.

[2] Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi and A. Mehmood, "Duplicate Questions Pair Detection Using Siamese MaLSTM," in IEEE Access, vol. 8, pp. 21932-21942, 2020.

[3] Priyanka Pachpande, Sharvari Govilkar, "Survey of various techniques for Insincere Question Classification", International journal of creative research thoughts, Refereed Journal, Vol. 9, January 2021.

[4] Sharma, Lakshay, Laura Graesser, Nikita Nangia and Utku Evci, "Natural Language Understanding with the Quora Question Pairs Dataset," ArXiv abs/1907.01041, 2019.

[5] Vyshnav MT, Sachin Kumar S, KP Soman, Offensive Language Detection: A Comparison, arXiv:2001.03131 [cs.CL], 2020.

[6] Sreelakshmi K, Premjith B, Soman K P, "Hate speech detection for Hindi-English Code-mixed social media text", in The Seventh International Symposium on Women in Computing and Informatics (WCI'19), 2019.

[7] G. Devi, M. Kumar, A., and Soman K. P., "Extraction of Named Entities from Social Media Text in Tamil Language Using N-Gram Embedding for Disaster Management", in Nature-Inspired Computation in Data Mining and Machine Learning, X. - S. Yang and He, X. - S., Eds. Cham: Springer International Publishing, pp. 207–223, 2020.

[8] C. Chen, "Insincere Question Classification by Deep Neural Networks," IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), pp. 15-23, 2021.

[9] Vani K and Deepa Gupta, "Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges," Information Processing & Management, 2018, Vol 54, Issue 3, pp. 408-432.

[10] Akshay Nautiyal and Deepa Gupta, "KCC QA Latent Semantic Representation using Deep Learning & Hierarchical Semantic cluster Inferential Framework," Procedia Computer Science, 2020 , 171, pp. 263–272.

[11] S. S. Teja Gontumukkala, Y. S. Varun Godavarthi, B. R. Ravi Teja Gonugunta, R. Subramani and K. Murali, "Analysis of Image Classification using SVM," 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 01-06, 2021.

[12] H. T. Le, D. T. Cao, T. H. Bui, L. T. Luong and H. Q. Nguyen, "Improve Quora Question Pair Dataset for Question Similarity Task," RIVF International Conference on Computing and Communication Technologies (RIVF), pp. 1-5, 2021.

[13] D. Y. Kim, X. Li, S. Wang, Y. Zhuo and R. K. -W. Lee, "Topic Enhanced Word Embedding for Toxic Content Detection in Q&A Sites," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1064-1071, 2019.

[14] Hao Mao , Rekha Kumar , Jerry Chen ,"Classification of Insincere Questions on Quora," 2019.

[15] Akshay Mungekar, Nikita Parab ,Prateek Nima, Sanchit Pereira, "Quora Insincere Questions Classification," National College of Ireland, 2019.

[16] R. Kumar, A. Kumar, M. Gupta and B. Chauhan, "Quora Based Insincere Content Classification & Detection for Social Media using Machine Learning," 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 294-299, 2021.

[17] H. Priyambowo and M. Adriani, "Insincere Question Classification on Question Answering Forum," International Conference on Electrical Engineering and Informatics (ICEEI), pp. 390-394, 2019.

[18] M. Chandra, A. Rodrigues and J. George, "An Enhanced Deep Learning Model for Duplicate Question Detection on Quora Question pairs using Siamese LSTM," IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), pp. 1-5, 2022.

[19] Pankajakshan V and Sridevi M, "Detecting Duplicate Question Pairs Using GloVe Embeddings and Similarity Measures," International Conference on Automation, Signal Processing, Instrumentation and Control, vol 700, pp 695–702, 2020.

[20] Zhang X, Sun X, and Wang H, "Duplicate Question Identification by Integrating FrameNet With Neural Networks," Thirty-Second AAAI Conference on Artificial IntelligenceProceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[21] H. Zhang and L. Chen, "Duplicate Question Detection based on Neural Networks and Multi-head Attention," International Conference on Asian Language Processing (IALP), pp. 13-18, 2019.

[22] Chunamari A, Yashas M, Basu, A, Anirudh, D K, and Soumya, C S, "Quora question pairs using XG boost," Emerging Research in Computing, Information, Communication and Applications, pp 715–721, 2021.

[23] Shankar, Shashi Kant, "Identifying Quora question pairs having the same intent," 2017.

[24] Y. Safali, G. Nergız, E. Avaroğlu and E. Doğan, "Deep Learning Based Classification Using Academic Studies in Doc2Vec Model," International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1-5, 2019.

[25] Lütfi Kerem Senel, Timo Schick and Hinrich Schütze, "Evaluating Language Understanding Capabilities of NLP Models With Context-Definition Alignment," Center for Information and Language Processing (CIS), Mar. 2022.