Classification of Insincere Questions Using Deep Learning: Quora Dataset Case Study

Iram Aslam[1], M. Azam Zia[1(✉)], Imran Mumtaz[1], Qamar Nawaz[1], and M. Hashim[2]

[1] Department of Computer Science, University of Agriculture, Faisalabad, Pakistan
[2] Faisalabad Business School, National Textile University, Faisalabad, Pakistan

Abstract. In the recent few year internet has gained the attention of researchers from all over the world. Internet has become crucial part of people's lives rapidly. The major cause of this rapid fame has been highlighted as its ability to make the work simpler and easier. Question answering websites have gained popularity due the human need of getting answers of queries. Such forums may be intruded by adversaries or hackers to ruin the forum's reputation. The detection of questions from such rivalry is still a key challenge because of two major reasons: firstly, the rate of users may be affected by such kind of questions, and secondly such insincere questions may also be asked to harm some particular user. Due to these reasons this study has been done to purpose deep learning-based solution for insincere question classification. Deep learning has been used due to its marvelous results in text classification related to various fields. Dataset has been collected of Quora website from publicly available source Kaggle. The machine learning methods SVM (Support vecttor Machine), Logistic regression has been compared with deep learning approach LSTM neural network (Long short-term memory). The dataset has been preprocessed and then used to train the models. Models have been trained on datasets by using MATLAB tool. This study has been also done to consider the impact of using various feature engineering techniques with models. Extensive experiments have been performed and parameters of the model have been optimized. F1 score has been used to compare the model performance because the dataset was highly imbalanced. Our existential results have highlighted that LSTM outperforms in terms of F1 score. The findings of this research would be beneficial for such question asking forums.

Keywords: Machine learning · Deep learning · Classification · Neural network · Feature engineering · Word embedding

1 Introduction

In recent few years there has been a remarkable increment in online forum usage, where the users can send their queries to such forums so that they can be posted

J. Xu et al. (Eds.): ICMSEM 2021, LNDECT 78, pp. 137–149, 2021.
https://doi.org/10.1007/978-3-030-79203-9_12

on the forum and other users can review their queries answer them or share them across the world. However due to usage of online forums on such large scale the need of filtering the content on such forums also has been arisen [2,14,15,17]. The content posted on such forums can be obtained in form of plain text. If there is no check on such type of content disrespectful, unprofessional or belligerent words may be posted on the site. Due to this there is a massive need of filtering out such content [5,22,24]. Sentiment analysis can also be performed on such dataset which has been defined as a computational method for automated content classification [1,26,27]. Different online question answering forums have been developed yet such as stack overflow, yahoo answers, Chegg, course-work, Quora etc. There has been some immature text filtration system used by such sites which has highlighted the need of a better system for classification of offensive content. The detection of unprofessional content can be intimidating task because of unavailability of adequate contextual and syntactical structures [6]. Therefore, this research has been motivated by the key challenge highlighted above. Different studies have been done to study the impact of deep learning on misclassification of data. The Quora dataset purposed various challenges for researchers to develop model for insincere question classification. Precision of models developed is still a major challenge needs to be resolved. So, this study is inspired by the issues discussed above. The aim of our work is to keep the issue under consideration said above. The approaches used by such previous studies may be helpful to perform such classification.

Ye et al., 2009 [25] have accentuated that the classification of the travel reviews using ML algorithms was considered for the people over the world. The information provided by the search engines has not satisfied the users and some mechanism was required to classify the travel blogs. They have implemented three ML algorithms SVM, NB and n-gram. Their study uncovered that SVM and n-gram provided superior results over NB.

Ghiassi et al., 2013 [12] divulged that traditional feature engineering techniques have used various features and techniques related to traditional text classification problems. They have introduced an n-gram based feature engineering method. Their method has provided better coverage of the twitter dataset and more accurate results. The results have been compared with traditional results by using an SVM based model. The dan2 model with n-gram ex-tracted features outperformed for twitter sentiment analysis.

Uyasal and Gunal [23] enunciated the impact of preprocessing techniques on the data.

The study have examined the two datasets of Turkish and English language with all preprocessing methods. Preprocessing have shown that it has positive impact regarding various dimensions related to text classification and analysis. It has been unveiled that instead of selecting some selective preprocessing techniques and overlooking other has not been correct choice. The selection of best combination of preprocessing methods which has to be suitable for the dataset has been considered the best solution.

Kewen et al., 2016 [8] reported that weighting the feature vectors has direct impact on classification accuracy of the model. Traditionally used tf/idf technique has been considered non suitable for the classification problems. To overcome such issues various weighting schemas have been purposed to overcome the flaws of traditional.

Zia et al., 2017 [28] unveiled that datamining has been used for tracking down the emanating research forums. There has been unearthed five features for efficiently extracting the emerging forums. Basically, four basic ML algorithms have been studied for this prediction. These models have been designed for examining the effect of purposed five features for spotlighting the rising research forums with ongoing upgradation.

Qaiser and Ali., 2018 [20] emphasized that tf/idf has a major shortcoming that is its inability to detect text with a little change. It has been seen that in such case unexpected results have been provided due considering words differently. It has been also pin pointed that this algorithm has no understanding of the semantic meanings which may lead to erroneous results. The word co-occurrence has been also overlooked in such cases.

Hartmann et al., 2019 [16] stated that machine learning has achieved so much fame and value in past few years. There has been a recent reawakening situation in this field as the developers are now capable of implementing more lucent algorithms. Many tasks have been made possible for researcher by this field of artificial intelligence and such tasks were considered tire-some in past. Due to which this field have so much potential and influencing capability. This branch of AI has been totally developed under the basis of mathematics and statistics. Just like AI it has been helpful for prediction of future events. Machine learning has been utilized in various domains such as sentiment analysis, weather prediction, text classification, face detection etc. But in case of big data there have been issues in feature engineering and text classification due to which the focus of researchers has been diverted from machine learning to deep learning.

Ahmad et al., 2019 [3] spotlighted that deep learning enabled the researchers to overcome the issues being faced due to shortcomings of machine learning. The algorithms of deep learning have been implemented using different feature engineering techniques from machine learning. In ML feature engineering techniques such as bag of words, bag of n grams, term frequency/inverse document frequency has been used while in deep learning pretrained or manually trained word embeddings have been used. There have been introduced different stages of the DL models and each stage has been designed to perform a specific task. In machine learning data has to be converted into structured form before training the model but in deep learning this is done by the layers of model.

Ain et al., 2020 [4] highlighted the usage of ML algorithms for detection of fraudulent queries. Various algorithms have been implemented on their preprocessed secondary dataset. Results of the classifiers have been analyzed and compared. Which revealed that Knn has outperformed in their case.

Sadiq et al., 2021 [21] asseverated that in-stead of manual feature engineering moving toward the neural networks has been considered as smart move.

The features have been automatically extracted through the layers of the network. The networks have been trained to learn the classification of text. The increment in the layers and neurons of the network has increased the model accuracy. The bi directional lstm and simple lstm has been implemented to learn the long-term dependencies between the documents of corpus. The nu-eral networks have been used for drastically improving the classification accuracy. The deep neural networks have been studied to improve the results and parameter optimization has been also considered. This research article has discussed the problems faced by the Quora website and purposed a deep learning model for question classification. This work is basically done for detection of insincere questions from Quora website dataset by implementing the deep learning and machine learning based models for evaluation and comparison of the accuracy of purposed model with ML models.

Contributions of the Purposed Study

Firstly, the dataset has been downloaded from Kaggle website. Then dataset has been refined and preprocessed to train the models.

Secondly the architecture of models has been designed and implemented in MATLAB tool which provides a separate module of deep learning and machine learn-ing of preprocessing data, training the model, displaying results and optimization of model parameters. The purposed research work includes the Support vector machine and logistic regression with three feature engineering techniques and a deep LSTM model with pretrained word embedding.

Thirdly extensive experiments have been performed to accuracy of the models is recorded and compared. Our experimental analysis enlightens that deep last outperforms in terms of F1 score as accuracy measurement, 82.5% then the other models.

2 Various Classifiers and Feature Engineering Techniques

Deep learning algorithms and methodologies have been stated as classifiers used for classification [3]. In this section, we have discussed the three algorithms which we have trained with our dataset. Each of the algorithm has been trained with three different feature engineer-ing techniques. The logistic regression and support vector machine algorithms have been trained with bag of words, term frequency/inverse document frequency and bag of n-grams model. Further the long short term memory network has been trained with three variants namely simple lstm, bilstm and deep lstm. Algorithms have been designed to classify the data using some mathematical or statistical function where the dataset for training the classifier is already available. We have used the Quora dataset available at Kaggle public ally for training the models discussed below and their variants.

2.1 Logistic Regression

A massive amount of data saved and streamed online cannot be filtered efficiently for extraction of useful information. Lack of valuable information results ineffective decision making [11]. Sentiment analysis has been defined as procedure

or methodology in which the mindset, point of view, opinions and attitude of the people has been assessed. Supervised and unsupervised learning algorithms have been used for big data classification. Various kinds of feature engineering techniques have been combined with LR such as bag of words, bag of ngrams and term frequency inverse document frequency [18].

2.2 Support Vector Machine (SVM)

Hyperplane boundaries have been used by linear classifiers and SVM has been classified into category of linear algorithms. SVM has been designed to work on the principle of mathematical function which maps the data of low dimensional data into high dimensional space. This mapping has been done to make the data linearly separable [7]. SVM algorithm has been designed to find out the optimal hyperplanes which has been used for separation of data points. Data points which have fallen on either side of the hyperplane can be classified into one of class. Hyperplanes has been considered to depend on the number of features in data set. Support vectors have been defined as the data points which have minimal distance from hyperplane. Support vectors have been taken into account because the orientation of the hyperplane has been affected by them [9,13].

2.3 LSTM (Long Short-Term Memory) Neural Network

LSTM has been introduced to overcome the inability of recurrent neural network. RNN has shortcoming that long term dependencies were not learned by this algorithm therefor LSTM has been introduced [10]. This algorithm has been designed such beautifully that learning the text sequence for long time has been its default behavior. LSTM has been designed with a special module addition called repeating module formed of four neural network layers. LSTM has been designed with a cell state which has the ability to add or remove the information when required [19].

3 Research Methodology

Firstly, the dataset has been obtained from the Kaggle website. The original dataset has three attributes qid, question-text and target. The dataset attribute target has been converted from numerical form 0 and 1 to categorical form sincere and insincere. The data set was highly imbalanced due to which the models were unable to learn the classification. Data set has been oversampled to unable the models to learn the classification. The Fig. 1 represents the data set distribution. Dataset has been distributed into train and test data using the 20/80 rule. Then different models are designed using MATLAB tool. These models are SVM+BOW, SVM+BON, SVM+TFIDF, LR+BOW, LR+BON, LR+TFIDF and LSTM+ Glove pretrained word embedding. The accuracy of all models is recorded and then compared to evaluate the model performance. Confusion matrix has been used performance measurement because our data set is highly

imbalanced and using the accuracy as performance measurement leads to the inefficient comparisons. This is because two models may have same accuracy value in case of imbalanced dataset. The dataset covers more than two hundred thousand records. The insincere questions have been identified on the basis of parameters like non-neutral tone, aggressive phrases, non-reality based etc. The data set is in textual form so it is converted in to numerical vectors for training by using the bag of words, bag of n-grams, ft/idf and glove feature engineering techniques. In the Fig. 2 the overall architecture of the methodology has been visualized. Nine classification models are being designed in developed in this study based on SVM, LR and LSTM. The eighty percent of the data is used to train the models and 20% of the data is used to test the models. In LSTM pretrained word embedding google vector is used to map our dataset to numerical vectors. Topic modeling is used as a powerful technique for analyzing the relationship between the data and for extraction of valuable patterns of dataset. Dataset is being analyzed by using LDA (Latent Dirichlet Allocation) to uncover the underlying topics in our dataset. The Fig. 3 shows the underlying four major topics in our dataset.

Figure 4 depicts the working of architecture of the purposed LSTM neural network of this study. Firstly, the dataset has been preprocessed and tokenized then using the glove pretrained word embedding tokenized documents are mapped to numerical vectors. A matrix of the dimension vocabulary size $\times$ document size is generated and fed as input into input layer. In dropout layer with rate of 50% dropout active neuron of input layer are turned off or dropped. Then in last layer performs the four gate operations and forward the output to SoftMax layer. The output is converted into probability by soft ax function and sincere or insincere output neuron is activated based on this probability value.

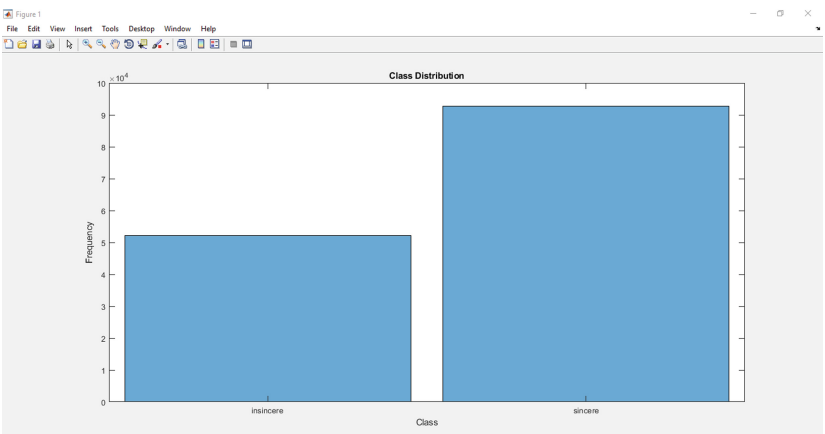

Fig. 1. Class distribution

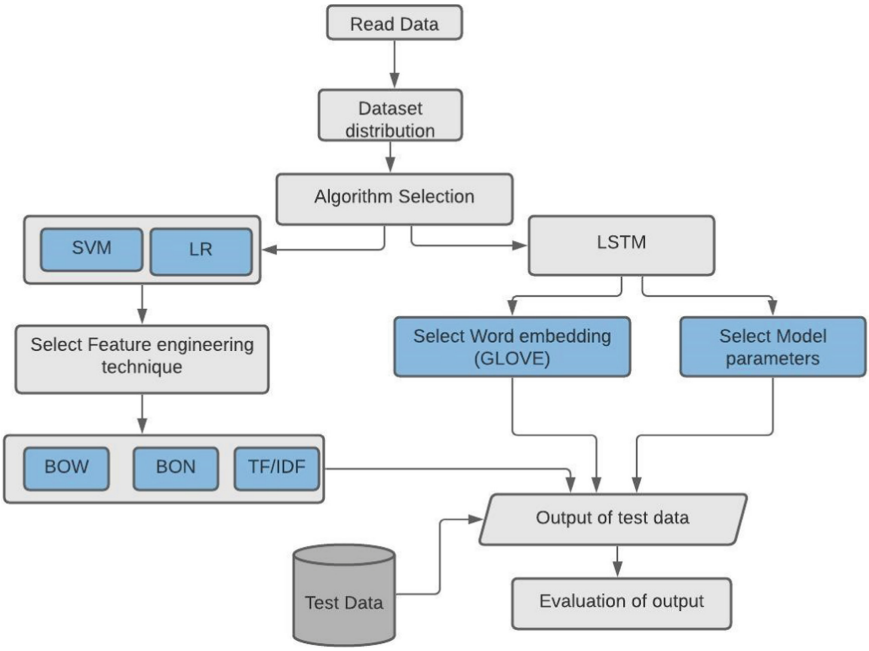

Fig. 2. Research architecture

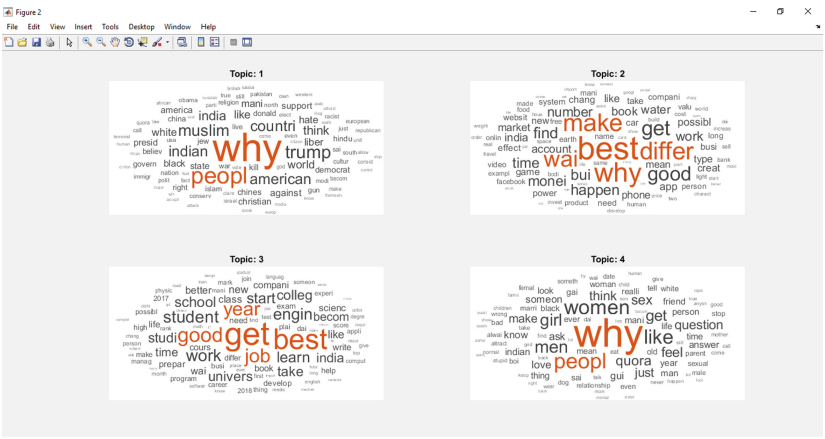

Fig. 3. LDA topic modeling

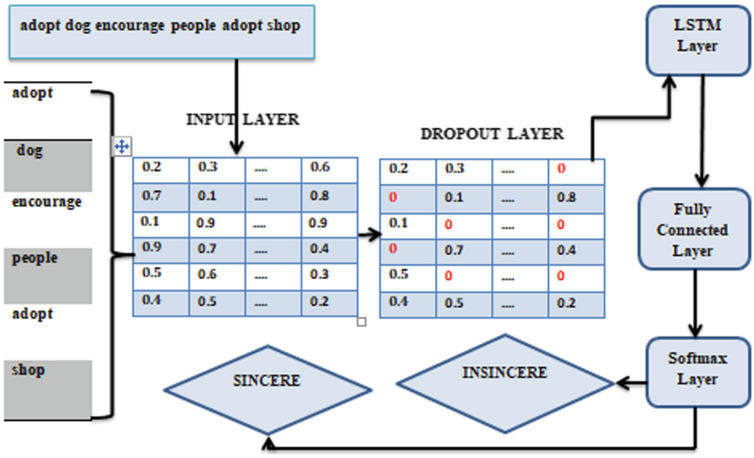

Fig. 4. Purposed LSTM architecture

4 Experimental Results

Experimental part of this study is discussed in this section. Various experiments per-formed based on accuracy. Achieved results are compared based on measurements used for accuracy and performance such as F1 score, true positive rate, false positive rate, true negative and false negative rate, precision recall and accuracy etc. These measurements for analyzing the model performance are computed with help of MATLAB tool. In Table 1 performance measurements of various feature engineering techniques are compared. From table it can be observed that SVM with bag of words outperforms in term of F1 score. LR with bag of n-grams is best in terms of lowest true negative rate.

Table 1. Performance measurement of feature engineering techniques with SVM & LR

Performance	SVM+BOW	SVM+BON	SVM+TF/IDF	LR+BOW	LR+BON	LR+TF\IDF
Accuracy	85.00%	83.00%	75.00%	82.00%	66.00%	81.00%
Error rate	15.00%	17.00%	25.00%	18.00%	24.00%	19.00%
TP rate	24.60%	22.80%	12.40%	22.90%	3.70%	21.00%
FP rate	60.90%	59.60%	63.00%	59.30%	63.30%	60.20%
TN rate	3.10%	4.40%	1.00%	4.80%	0.70%	3.80%
FN rate	11.30%	13.10%	23.60%	13.00%	32.30%	15.00%
Precision	88.00%	83.00%	92.00%	82.00%	83.00%	84.00%
Recall	68.00%	63.00%	34.00%	63.00%	10.00%	58.00%
F1-Score	78.00%	73.00%	63.00%	72.00%	46.00%	71.00%

All performance measure weighted averages are recorded using MATLAB tool. TP rate shows the insincere questions being classified correctly while false positive rate envisages correctly classified insincere questions. F1 measure is used as a balancing factor be-cause of imbalanced dataset The F1 measure is most suitable for balancing the output. Precision and Recall are calculated to determine the behavior of the algorithms. After implementing the ML algorithms, purposed LSTM architecture is developed in MATLAB. Experiments are performed with architecture such as single layer architecture, bi-directional list and deep lstm are trained and then performance is evaluated.

Table 2 depicts the parameter settings used to perform these experiments. In the all three models pretrained word embedding google vectors is used to train our models. In LSTM single layer architecture number of layers is 6. Then 7, 10 are respectively in the bi-directional architecture and deep LSTM architecture. The number of the hidden units used in these three architectures is 180,280 and 430 respectively. The size of input vector used in three models is 300x25. Dropout rate to turnoff active neurons is set to 50% in initial two models and 20% in deep architecture of LSTM. In term of specificity SVM with bag of words model outperforms the other architecture. LR with bag of n gram model has the highest error rate among all architectures of our models.

Table 2. LSTM model parameters

Parameters	LSTM	Bi-LSTM	Deep LSTM
Word embedding	Glove	Glove	Glove
Layers	6	7	10
Hidden units	180	280	430
Input dimension	300×25	300×25	300×25
Dropout rate	50%	50%	20%
Output size	180	180	180

In the Fig. 5 the single layer last, Fig. 6 bilist training and in Fig. 7 the deep lstm training progresses are envisaged. Purposed LSTM model performs better then ML but extensive experiments shows that increment in neurons and layers increase the overall model performance and decreases the loss rate. Bi-LSTM performance in comparison of the other two variants is minimum. The Table 3 shows the LSTM and its variant performance analysis which shows that deep lstm outperforms other models trained. By analyzing the results envisaged in this table we conclude that deep architecture of our models takes highest time among LSTM, bi-LSTM and deep LSTM. The loss rate is minimum in deep

LSTM and same in the other two models in contrast with it. When we observe accuracy rate, we observe that LSTM and deep LSTM have same accuracy level while addition of bi-directional layer in our model decreased the model accuracy. When we observe the F1 score of LSTMs and deep LSTM it is concluded that former has higher accuracy level.

Table 3. Performance analysis of LSTM and variants

Performance	LSTM	Bi-LSTM	Deep LSTM
Training time (m)	4148	6303	8014
Iterations per epoch	7238	7238	7238
Loss	0.6	0.6	0.5
Accuracy (%)	87	85	87
Error rate (%)	13	15	13
TP rate (%)	27.9	24.6	28.4
FP rate (%)	4.4	3.1	4.9
TN rate (%)	59.6	60.9	59.1
FN rate (%)	8	11.3	7.5
Precision (%)	85	88	88
Recall (%)	79	68	77
F1-Score (%)	82	78	82.5

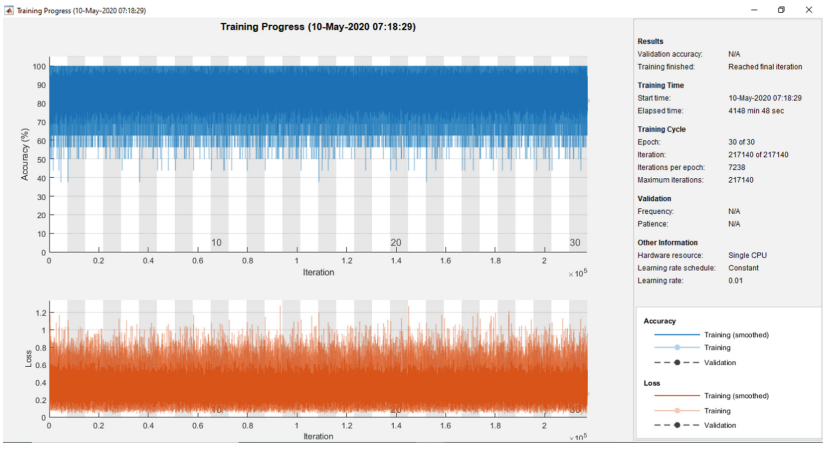

Fig. 5. LSTM training progress

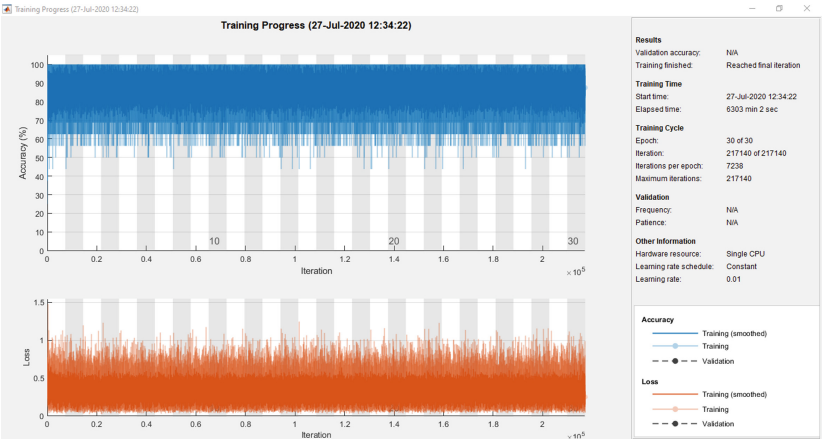

Fig. 6. Bi-LSTM training progress

Fig. 7. Deep LSTM training progress

5 Conclusion

With tremendous increase in use of internet vast range of problem is also occurring. One of these problems addressed in this study is classification of insincere content from question answering forums. Classification of insincere questions of Quora using ML and DL is addressed in this research work which is overlooked in previous works. Quora dataset available public ally on Kaggle website is downloaded for this study. Experiments are performed and the performance of each model is recorded. If accuracy is used as performance measurement SVM+BOW obtained 85% accuracy which highest in ML models analyzed while deep LSTM in DL has obtained highest accuracy rate which is 87%. But due to imbalance dataset accuracy of each model is compared in terms if F1 score. Critical analysis

of the model is done in this research work. SVM with bag of words has obtained the highest F1 score which is 78% while in variants of LSTM the deep LSTM obtained the highest F1 score which is 82.5%. Deep LSTM also minimized the loss value. Accuracy of the deep LSTM model can still be improved.

6 Significance of Study

The findings of this research work will be helpful for question answering forums to protect their sites and users from adversaries by filtering out the insincere content. This work will be also encouraging for working on such dataset and deep leaning algorithms further.

7 Future Study

In future, we would like to extend our work by focusing towards a deeper neural network and also implementing other neural networks such convolutional neural network, capsule networks etc. We will try to develop an automated insincere content classifier for question answering forums.

Acknowledgement. I feel much honor to express indebtedness to my honorable supervisor and I would like to thanks Dr. M. Azam Zia for encouraging and supporting me. It wishes to record my sincere appreciation toward him for his mo-tivation behind this study. I would also like to thank my parents for their support.

References

1. Aborisade, O., Anwar, M.: Classification for authorship of tweets by comparing logistic regression and Naive Bayes classifiers. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 269–276. IEEE (2018)
2. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 163–222. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_6
3. Ahmad, S., Asghar, M.Z., et al.: Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. HCIS **9**(1), 24 (2019)
4. Ul Ain, Q., Zia, M.A., Asghar, N., Saleem, A.: Analysis of variant data mining methods for depiction of fraud. In: Xu, J., Duca, G., Ahmed, S.E., García Márquez, F.P., Hajiyev, A. (eds.) ICMSEM 2020. AISC, vol. 1190, pp. 423–432. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49829-0_31
5. Alasadi, S.A., Bhaya, W.S.: Review of data preprocessing techniques in data mining. J. Eng. Appl. Sci. **12**(16), 4102–4107 (2017)
6. Alhaj, Y.A., Xiang, J., et al.: A study of the effects of stemming strategies on Arabic document classification. IEEE Access **7**, 32664–32671 (2019)
7. Chatterjee, S., Jose, P.G., Datta, D.: Text classification using SVM enhanced by multithreading and Cuda. Int. J. Mod. Educ. Comput. Sci. **11**(1) (2019)
8. Chen, K., Zhang, Z., et al.: Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Syst. Appl. **66**, 245–260 (2016)

9. Elnagar, A., Al-Debsi, R., Einea, O.: Arabic text classification using deep learning models. Inf. Process. Manag. **57**(1), 102121 (2020)
10. Eryılmaz, E.E., Şahin, D.Ö., Kılıç, E.: Filtering Turkish spam using LSTM from deep learning techniques. In: 2020 8th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–6. IEEE (2020)
11. Genkin, A., Lewis, D.D., Madigan, D.: Large-scale Bayesian logistic regression for text categorization. Technometrics **49**(3), 291–304 (2007)
12. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network. Expert Syst. Appl. **40**(16), 6266–6282 (2013)
13. Ghosh, S., Dasgupta, A., Swetapadma, A.: A study on support vector machine based linear and non-linear pattern classification. In: 2019 International Conference on Intelligent Sustainable Systems (ICISS), pp. 24–28. IEEE (2019)
14. Hao, F., Min, G., et al.: k-clique community detection in social networks based on formal concept analysis. IEEE Syst. J. **11**(1), 250–259 (2015)
15. Hao, F., Park, D.S., Pei, Z.: When social computing meets soft computing: opportunities and insights. HCIS **8**(1), 1–18 (2018)
16. Hartmann, J., Huppertz, J., et al.: Comparing automated text classification methods. Int. J. Res. Mark. **36**(1), 20–38 (2019)
17. Hassan, A., Mahmood, A.: Deep learning for sentence classification. In: 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1–5. IEEE (2017)
18. Khanday, A.M.U.D., Rabani, S.T., et al.: Machine learning based approaches for detecting COVID-19 using clinical text data. Int. J. Inf. Technol. **12**(3), 731–739 (2020)
19. Murthy, D., Allu, S., et al.: Text based sentiment analysis using LSTM. Int. J. Eng. Res. Tech. Res. **9**(05) (2020)
20. Qaiser, S., Ali, R.: Text mining: use of TF-IDF to examine the relevance of words to documents. Int. J. Compu. Appl. **181**(1), 25–29 (2018)
21. Sadiq, S., Mehmood, A., et al.: Aggression detection through deep neural model on twitter. Futur. Gener. Comput. Syst. **114**, 120–129 (2021)
22. Sriram, B., Fuhry, D., et al.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841–842 (2010)
23. Uysal, A.K., Gunal, S.: The impact of preprocessing on text classification. Inf. Process. Manag. **50**(1), 104–112 (2014)
24. Yan, D., Guo, S.: Leveraging contextual sentences for text classification by using a neural attention model. Comput. Intell. Neurosci. (2019)
25. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. **36**(3), 6527–6535 (2009)
26. Zhang, Z., Zou, Y., Gan, C.: Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression. Neurocomputing **275**, 1407–1415 (2018)
27. Zhou, W., Wang, H., et al.: A method of short text representation based on the feature probability embedded vector. Sensors **19**(17), 3728 (2019)
28. Zia, M.A., Zhang, Z., et al.: Prediction of rising venues in citation networks. J. Adv. Comput. Intell. Intell. Inf. **21**, 650–658 (2017)