

# Supervised Classification Based Machine Translation Quality Estimation

Nivedita Bharti, Nisheeth Joshi, Iti Mathur

**Abstract:** *This submission describes the study of linguistically motivated features to estimate the translated sentence quality at sentence level on English-Hindi language pair. Several classification algorithms are employed to build the Quality Estimation (QE) models using the extracted features. We used source language text and the MT output to extract these features. Experiments show that our proposed approach is robust and producing competitive results for the DT based QE model on neural machine translation system.*

**Keywords:** *Machine Translation, Quality Estimation, Classification Algorithms, Features, Performance Evaluation.*

## I. INTRODUCTION

Quality Estimation (QE) is defined as a problem automatically evaluating the translation output without using the reference translations [1-3]. There are several engrossing applications of this QE problem: first, to tell the readers of the target language about whether they can trust on the generated translation output or not. Second, determining whether the obtained translation is adequate for publishing as it is. Third, choosing the best translation generated from the several MT systems. Last but not the least, filtering out the translated language sentences that are of poor quality for the purpose of post-editing by the proficient translators.

With this submission we tried to address the problem as predicting the translation quality on sentence level as a discretized classification task to a single translated output corresponding to a given source sentence. In other words, using the given source language sentence and its generated translation output for feature extraction, the developed QE model is asked to label the quality to the translated sentence as Excellent or Good, or Average or Poor.

In section II we write about the works done by various researchers in the past. Section III present about the proposed methodology, and in section IV, we briefly explained the various evaluation metrics employed for analyzing the performance of the developed quality estimation models. Section V presented the results obtained by our models. Finally, section VI concluded our work.

**Revised Manuscript Received on February 19, 2020.**

\* Correspondence Author

**Nivedita Bharti\***, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. E-mail: [nivedita2bharti@gmail.com](mailto:nivedita2bharti@gmail.com)

**Nisheeth Joshi**, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. E-mail: [nisheeth.joshi@rediffmail.com](mailto:nisheeth.joshi@rediffmail.com)

**Iti Mathur**, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. E-mail: [mathur\\_iti@rediffmail.com](mailto:mathur_iti@rediffmail.com)

## II. RELATED WORKS

The first comprehensive study on sentence level QE was done by Blatz et al. [1]. The authors trained various classifiers and regressors on the features extracted for the translation outputs labeled using the NIST MT evaluation metric [4]. Further, Quirk [5] exploited classifiers together with a pre-defined threshold for “GOOD” and “BAD” translation outputs. For this problem, they used a small size of 350 translation outputs whose quality is manually labeled. Gamon et al. [6] developed a classifier-based QE model with the linguistic features. These features were extracted from the human and machine translations for finally differentiating these two types of translations using SVM classifier built with linear and non-linear combinations of LM. However, the predictions showed very low correlations with manual judgments. Albrecht & Hwa [7] trained a regression-based QE model using the features extracted from the pseudo-references and translation output. Although, the approach used by the authors is remained necessarily a reference-dependent translation evaluation setting, because alternative MT Engines are compulsory in this scenario. Pado et al. [8] used regression-based approach to the quality estimation model built with the features using the textual entailment between the translated sentence and the reference translation.

Work by Xiong et al. [9] was focused on detecting the word error through WER estimation. To implement this approach the authors used neighboring words POS tags together with the link grammar type of parser. Specia et al. [10] exploited linguistic motivated features like chunking, POS tagging, named entities and dependency relations to estimate the translation quality of English-Arabic language pair. Hardmeier [11] employed the dependency and constituency structures for estimating the translation quality of English-Swedish and English-Spanish language pairs. Avramidis [12] performed the automatic ranking of multiple translation outputs at sentence level by exploiting the adequacy information like nouns, verbs, subordinate clauses, sentences and punctuation occurrences as the features. Mehdad et al. [13] conducted the adequacy estimation of translation by using the cross-lingual textual entailment by pushing the semantics in evaluating the translation output without exploiting the reference translations. Bojar et al. [14] exploited other translation systems for translating the source sentence into target sentence and vice-versa.

After this, to estimate the sentence level translation quality, they compared the results with original source and target language sentences. Kreutzer et al. [15] developed deep neural based feed-forward neural network model for word level translation quality prediction. For this problem, they processed the concatenation of neighbouring target language words and source language words vector embeddings into the feature vectors. Blain et al. [16] developed word and sentence level translation quality models by exploiting the bilinear embeddings for modeling the relationship strength of source language and target language words. Kim et al. [17, 18] trained a three-level stacked architecture of RNN based on neural MT [19] for the development of QE models at sentence level. To implement the proposed model, the authors combined a neural based word prediction model with the translation QE models of word and sentence level. Later, the above neural MT model was replaced by Wang et al. [20] with a modified version of self-attention-based transformer model [21] for estimating the English-German language-based sentence level translation quality estimation.

Hou et al. [22] did sentence level QE for English-to-German language pair. The authors developed two different techniques by exploiting two stage based neural quality assessment models. In which, the first stage consists of a feature extraction module and the second step includes a quality estimator. Particularly, one of the developed approaches called as BERT-based QE model used external monolingual knowledge of source and target language sentences both which were generated by pre-trained neural self-attention models. While the other one called as Bi-directional QE model used translation knowledge from two different translation directions between the two languages. Kim et al. [23] employed a novel approach for estimating the translation quality on sentence level. They used bilingual BERT based quality estimation exploiting the multi-task learning technique. Kepler et al. [24] combined neural, linear and predictor-estimator models together with the novel transfer learning techniques via exploiting the pre-trained models namely: XLM and BERT for developing the QE models at word, sentence and document levels all.

## III. PROPOSED METHODOLOGY

Our proposed work mainly focused on the implications of a wide range of linguistic features that represents the different aspects of the translation output quality. To perform the experiments of quality estimation on MT engines, a corpus was required onto which machine learning systems can be trained. For this task, we have considered the English-Hindi parallel corpus in tourism and health domain from TDIL, MeitY, GoI that was developed under the project, "Development of English to Indian Languages Machine Translation System". Thus, we have a total of 30,000 parallel sentences of English-Hindi in the corpus. Among this the total dataset is split into the following three sets:

- Training set: 70% for training the model
- Development set: 15% for optimizing the developed model
- Test set: 15% to test the developed model

Table I depicted the statistics of the parallel corpora, and Table II depicted the dataset split used for the supervised model.

**Table- I: Statistics of the parallel corpora**

S. No.	Number of Sentences	Domain
1	15000	Tourism
2	15000	Health

**Table- II: Dataset split for the model development**

Dataset	Number of Sentences
Total	30,000
Training set	21,000
Development set	4,500
Testing set	4,500

## A. Translation Systems

We have used three machine translation systems: Google, Bing, and Moses Phrase-based Model<sup>13</sup> to obtain the alternative candidate translations corresponding to a given source sentence. Among which, Google and Bing are a web based neural translator, and Moses-Phrase based is a statistical machine translator [25, 26].

## B. Features

An algorithm which extract features from this corpus was developed. We have used feature set which contain 27 features. The list of features which were used in training the QE classifiers is listed below in Table III. These features were extracted by analyzing the source and target language sentence. Based on these obtained features, the classes in the dataset was manually labeled. These classes are:

- Poor
- Average
- Good
- Excellent

These four classes depicted the translation of English sentences into Hindi by MT Engines. Once done, the classifiers were trained on this data.

**Table- III: Features used in training the QE model**

S. NO.	Features Description
1	n-gram probabilities (unigram, bigram and trigram) of input language sentence
2	n-gram probabilities (unigram, bigram and trigram) of output language sentence
3	Parts of Speech (POS) Tags of input language sentence
4	POS Tags of output language sentence
5	Translation Probabilities
6	Token counts of the input language sentence
7	Token counts of the output language sentence
8	Average token length of input language sentence

9	Average token length of output language sentence
10	Percentage of high and low frequency uni-gram, bi-gram, tri-gram of the input language sentence
11	Percentage of high and low frequency uni-gram, bi-gram, tri-gram of the output language sentence
12	Count of punctuation marks of the input language sentence
13	Count of punctuation marks of the output language sentence

### C. Classification Algorithms

The classifiers viz: Decision Trees [27], SVM [28] and MLP algorithm [29] were trained. In total three classifiers were developed for this problem. Finally, the performances of the classifiers-based QE models were evaluated with the standard evaluation metrics. Section 4 discusses the evaluation metrics in more detail. The proposed architecture of the developed QE model using supervised learning approach as depicted below in Fig. 1.

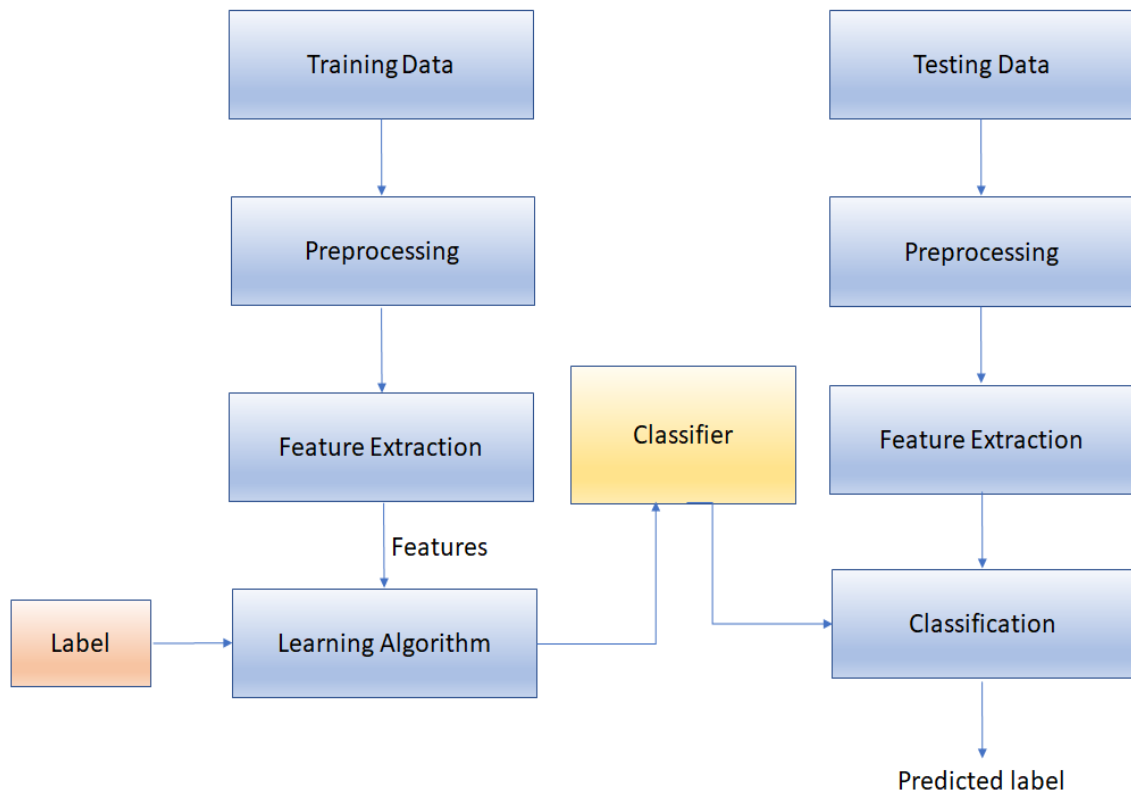


Fig. 1. Machine learning based approach of translation Quality Estimation.

### IV. PERFORMANCE EVALUATION METRICS

In any field of computer science, it is practically a common necessity that the performance of the developed machine learning based models has to be measured and compared with each other. In this context, to interpret the appropriateness of a developed translation QE models a simple performance evaluation mechanism that is commonly used in supervised learning classification approach is a confusion matrix (as shown in Table IV). It is basically used in order to depict the test output of a classifier model. In this matrix every column shows the values belong to the predicted category, while every row shows the values belongs to the actual category.

Table- IV: Confusion matrix.

Confusion Matrix		Predicted Category	
		Positive Category	Negative Category
Actual Category	Positive Category	TP	FN
	Negative Category	FP	TN

Using this confusion matrix, numerous evaluation metrics equations are obtained and these equations are very important for the performance evaluation. The important evaluation measures that are considered here are explained below:

## ➤ Precision

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

## ➤ Recall

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

## ➤ F-Measure

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

## ➤ Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

## V. RESULTS

This section shows the results of our developed QE models. Particularly, the performance of our developed sentence level DT based, SVM based and MLP based QE models on test set are represented below in Fig. 2, Fig. 3 and Fig. 4 respectively. The evaluation results revealed that the DT based QE model performs better than the SVM and MLP models with all evaluation metrics i.e. Precision, Recall, F-measure and Accuracy. Furthermore, all the three used QE models ranked MT system2 at top. Particularly, the best set up surprisingly achieved the same Precision, Recall, and F-measure value as 0.874, Accuracy=87.41%.

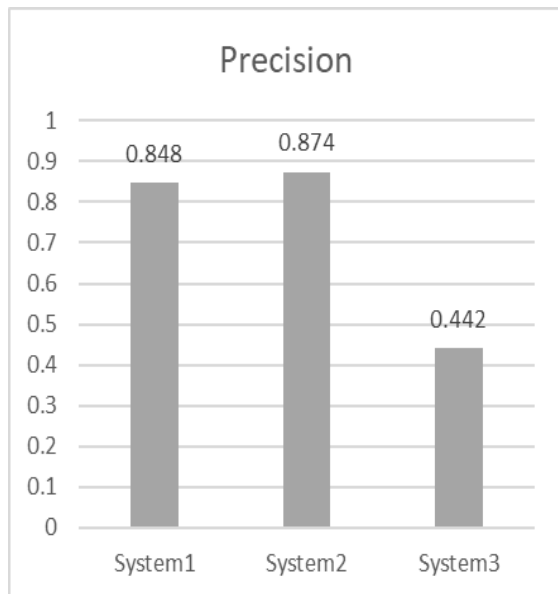


Fig. 2A

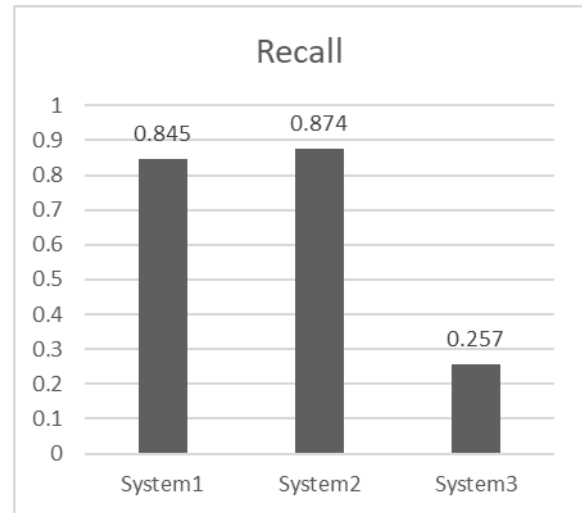


Fig. 2B

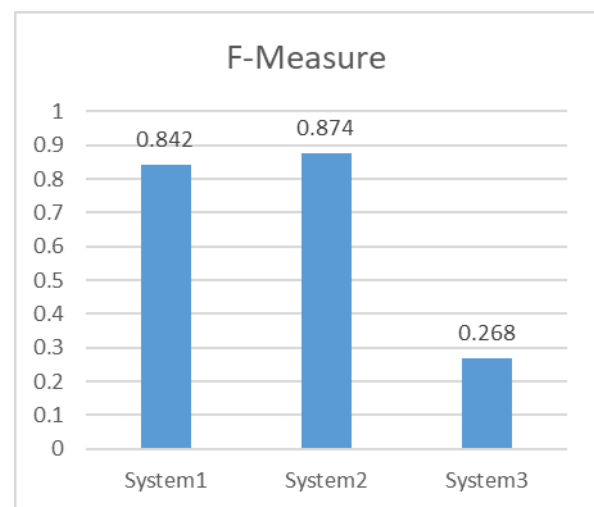


Fig. 2C

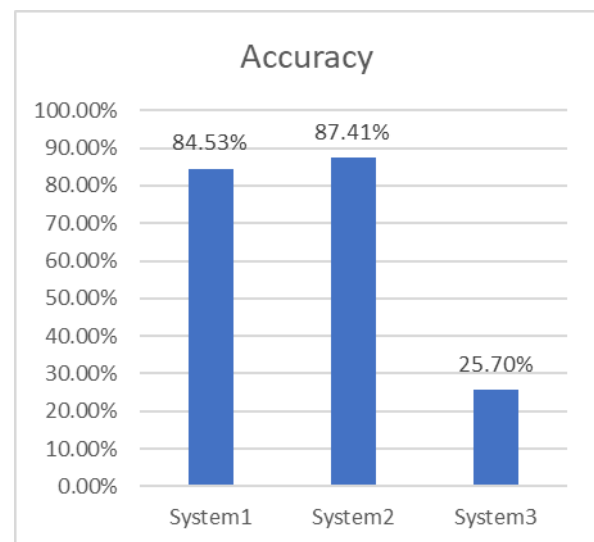


Fig. 2D

**Fig. 2: DT based QE model performance Evaluation: (2A) DT based QE model Precision (2B) DT based QE model Recall (2C) DT based QE model F-measure (2D) DT based QE model Accuracy.**

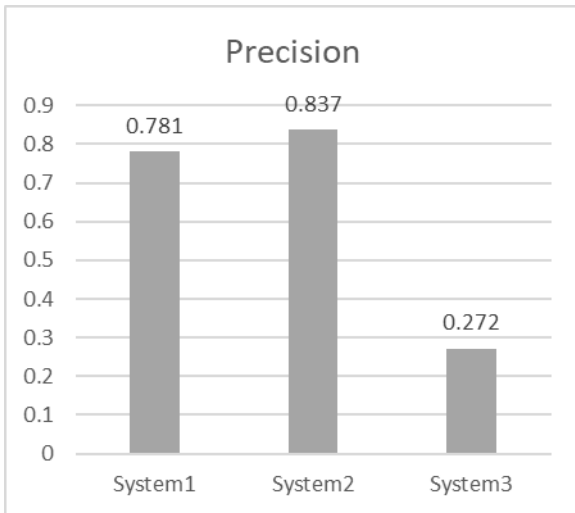


Fig. 3A

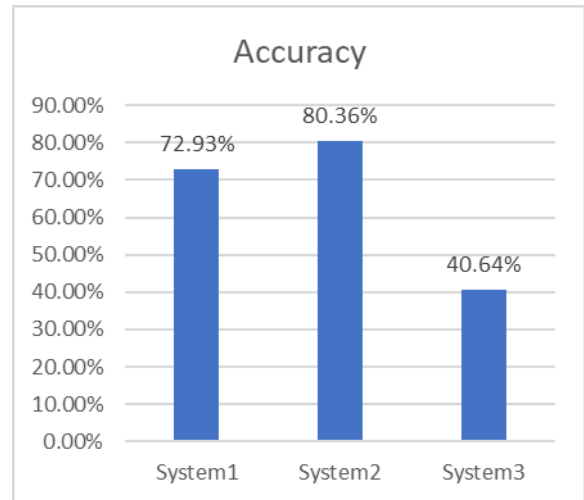


Fig. 3D

**Fig. 3: SVM based QE model performance Evaluation:**  
(3A) SVM based QE model Precision (3B) SVM based QE model Recall (3C) SVM based QE model F-measure (3D) SVM based QE model Accuracy.

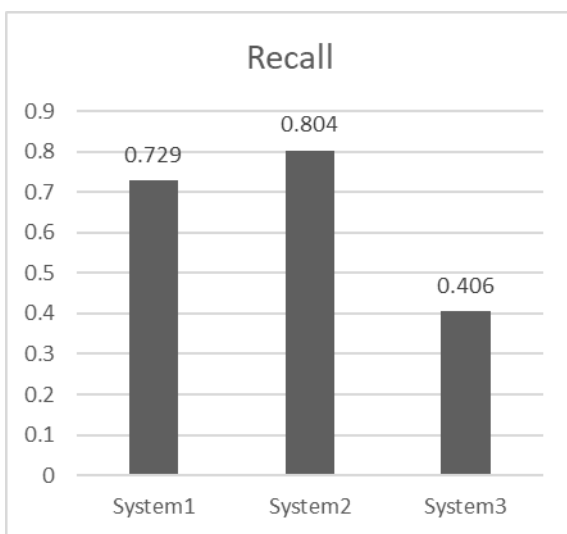


Fig. 3B

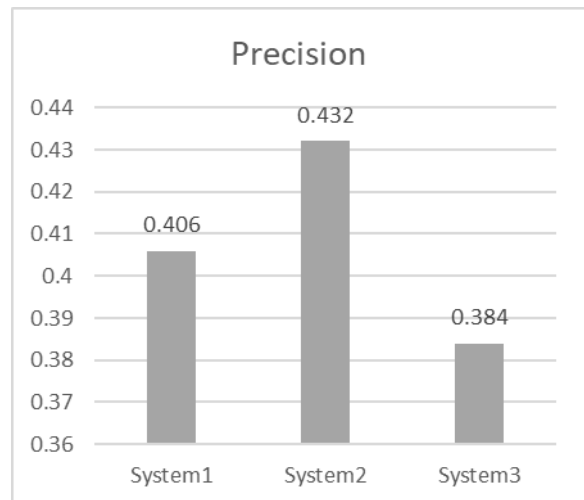


Fig. 4A

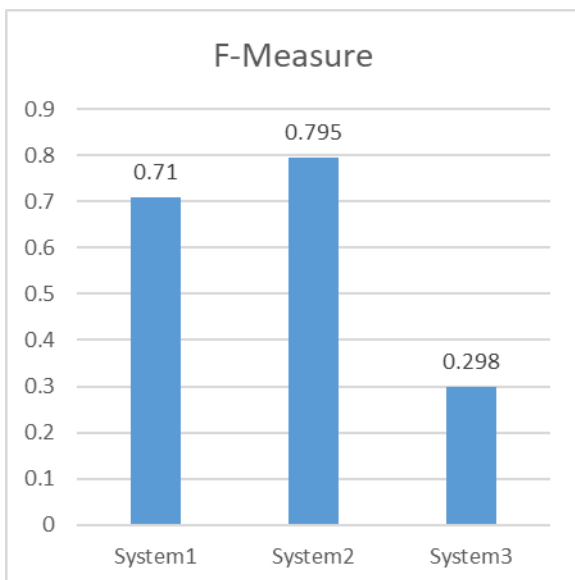


Fig. 3C

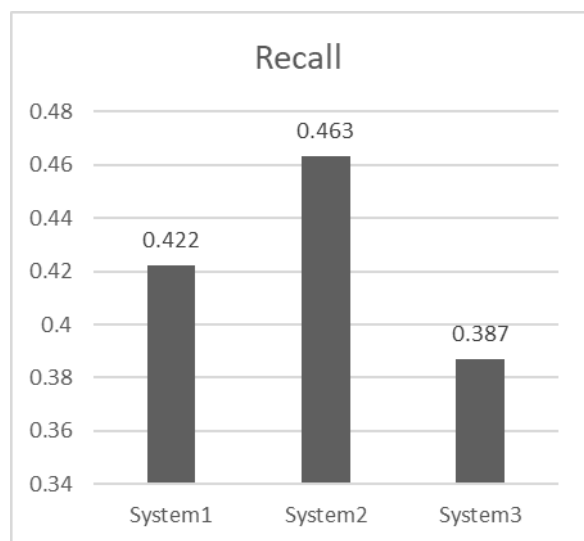


Fig. 4B



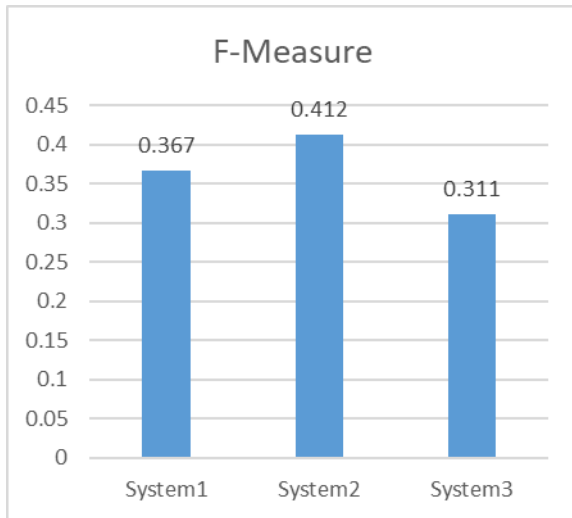


Fig. 4C

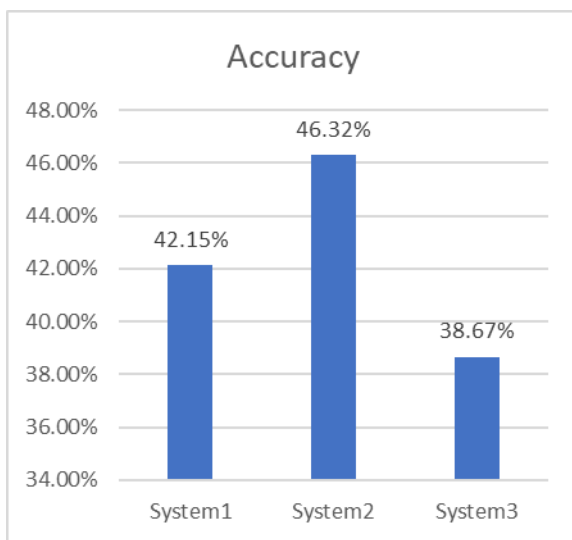


Fig. 4D

**Fig. 4: MLP based QE model performance Evaluation:**  
**(4A) MLP based QE model precision (4B) MLP based QE**  
**model Recall (4C) MLP based QE model F-measure (4D)**  
**MLP based QE model Accuracy.**

## VI. CONCLUSIONS

This paper reports our work on quality estimation task at sentence level on English-Hindi dataset. For this work, we proposed a supervised machine learning based QE models as a classification problem using different classifiers built with the extracted features set. Further, the developed models' performances were analyzed and compared with each other. Among the proposed methods, DT based QE models showed better results compared to other classification algorithms-based models. Furthermore, MT System2 showed best results among the three MT systems. Finally, the experimental results on unseen test set showed the effectiveness of our developed classification-based QE models even without referring the reference translations.

## REFERENCES

1. J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, N. Ueffing, "Confidence estimation for machine translation," *InProc. of the 20th international conference on computational linguistics*, pp. 315-321, 2004.

2. L. Specia, K. Shah, J.G. De Souza, T. Cohn, "QuEst-A translation quality estimation framework," *InProc. of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 79-84, 2013.
3. L. Specia, C. Scarton, G.H. Paetzold, "Quality estimation for machine translation," *Synthesis Lectures on Human Language Technologies*, vol. 11, no. 1, pp. 1-62, 2018.
4. G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," *InProc. of the second international conference on Human Language Technology Research*, pp. 138-145, 2002.
5. C. Quirk, "Training a Sentence-Level Machine Translation Confidence Measure," *InLREC*, pp. 825-828, 2004.
6. M. Gamon, A. Aue, M. Smets, "Sentence-level MT evaluation without reference translations: Beyond language modeling," *InProc. of EAMT*, pp. 103-111, 2005.
7. J. Albrecht, R. Hwa, "Regression for sentence-level MT evaluation with pseudo references," *InProc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 296-303, 2007.
8. S. Padó, M. Galley, D. Jurafsky, C.D. Manning, "Textual entailment features for machine translation evaluation," *InProc. of the Fourth Workshop on Statistical Machine Translation*, pp. 37-41, 2009.
9. D. Xiong, M. Zhang, H. Li, "Error detection for statistical machine translation using linguistic features," *InProc. of the 48th annual meeting of the Association for Computational Linguistics*, pp. 604-611, 2010.
10. L. Specia, N. Hajlaoui, C. Hallett, W. Aziz, "Predicting machine translation adequacy," *InMachine Translation Summit*, vol. 13, no. 2011, pp. 19-23, 2011.
11. C. Hardmeier, "Improving machine translation quality prediction with syntactic tree kernels," *InEAMT 2011*, pp. 233-240, 2011.
12. E. Avramidis, "Comparative quality estimation: Automatic sentence-level ranking of multiple machine translation outputs," *InProc. of COLING 2012*, pp. 115-132, 2012.
13. Y. Mehdad, M. Negri, M. Federico, "Match without a referee: evaluating MT adequacy without reference translations," *InProc. of the Seventh Workshop on Statistical Machine Translation*, pp. 171-180, 2012.
14. O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, "Findings of the 2014 workshop on statistical machine translation," *InProc. of the ninth workshop on statistical machine translation*, pp. 12-58, 2014.
15. J. Kreutzer, S. Schamoni, S. Riezler, "Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation," *InProc. of the Tenth Workshop on Statistical Machine Translation*, pp. 316-322, 2015.
16. F. Blain, C. Scarton, L. Specia, "Bilexical embeddings for quality estimation," *InProc. of the Second Conference on Machine Translation*, pp. 545-550, 2017.
17. H. Kim, H.Y. Jung, H. Kwon, J.H. Lee, S.H. Na, "Predictor-estimator: Neural quality estimation based on target word prediction for machine translation," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 17, no. 1, pp. 1-22, 2017a.
18. H. Kim, J.H. Lee, S.H. Na, "Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation," *InProc. of the Second Conference on Machine Translation*, 2017b, pp. 562-568.
19. D. Bahdanau, K. Cho, Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
20. J. Wang, K. Fan, B. Li, F. Zhou, B. Chen, Y. Shi, L. Si, "Alibaba submission for WMT18 quality estimation task," *InProc. of the Third Conference on Machine Translation: Shared Task Papers*, pp. 809-815, 2018.
21. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," *InAdvances in neural information processing systems*, pp. 5998-6008, 2017.
22. H. Qi, "NJU Submissions for the WMT19 Quality Estimation Shared Task," *InProc. of the Fourth Conference on Machine Translation: Shared Task Papers*, vol. 3, pp. 95-100, 2019.

23. H. Kim, J.H. Lim, H.K. Kim, S.H. Na, "QE BERT: Bilingual BERT using Multi-task Learning for Neural Quality Estimation," *InProc. of the Fourth Conference on Machine Translation: Shared Task Papers*, vol. 3, pp. 85-89, 2019.
24. F. Kepler, J. Trénous, M. Treviso, M. Vera, A. Góis, M.A. Farajian, A.V. Lopes, A.F. Martins, "Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task," *InProc. of the Fourth Conference on Machine Translation*, pp. 80–86, 2019.
25. P. Koehn, F.J. Och, D. Marcu, "Statistical phrase-based translation," *InProc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 1, pp. 48-54, 2003.
26. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, "Moses: Open source toolkit for statistical machine translation," *InProc. of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177-180, 2007.
27. L. Breiman, J.H. Friedman, R.A. Olshen, & C.J. Stone, "Classification and regression trees," Routledge, pp. 151-166, 1984.
28. V.N. Vapnik, "The nature of statistical learning," Theory, 1995.  
G.E. Hinton, "Connectionist learning procedures," *InMachine learning*, Morgan Kaufmann, pp. 555-610, 1990.