






Subjective Answer Evaluation Using NLP

Sheela S Maharajpet *¹, Navya D †², and Sonam Bhandurge ‡³

¹Assistant Professor, Department of MCA, Acharya Institute of Technology, Bangalore

²Department of MCA, Acharya Institute of Technology, Bangalore

³Assistant Professor, Angadi Institute of Technology & Management, Belagavi

Abstract

An overview of the state of NLP techniques for assessing subjective responses is given in this abstract. Semantic analysis, sentiment analysis, and coherence verification are the three main areas of emphasis. Understanding the meaning hidden within the text is the goal of semantic analysis, and it can be accomplished with tools like knowledge graphs, word embeddings, and transformer models (BERT, GPT, etc.). Sentiment analysis evaluates the response's subjective subtleties and emotional tone. The process of coherence checking guarantees the text's consistency and logical flow. Because of the inherent heterogeneity in human language and the variety of ways that various people may convey the same notion, evaluating subjective replies is a difficult undertaking. Due to the heavy reliance of traditional assessment systems on human evaluators, issues with bias, scalability, and consistency arise. A potential remedy is provided by natural language processing (NLP), which gives methods and tools for automating and standardizing the evaluation process.

Keywords: Sentiment analysis, Logical flow, Language, Evaluation.

*Email: sheelamaharajpet4@gmail.com Corresponding Author

†Email: navyad.22.mcav@acharya.ac.in

‡Email: sonambhandurge.aitm@gmail.com

1 Introduction

It is possible to conduct a thorough evaluation of pupils' performance using open-ended, subjective questions and answers that are based on individual viewpoints and conceptual understanding. Although there are no restrictions on the replies, subjective responses are very different from objective ones in terms of length, time commitment, and the necessity for more attention to detail and objectivity when grading because of their rich contextual content.

Due to natural language's inherent ambiguity, analysing subjective replies with computers is difficult. Before using several techniques to compare textual data, such as document similarity, latent semantic structures, idea networks, and ontologies, preprocessing activities like data purification and tokenization are essential.

This topic has been tackled in a number of ways, and this study explores potential avenues for future development. Because of their contextual character, subjective tests are sometimes viewed as more difficult by both professors and students. The laborious process of carefully examining each word for scoring, in addition to the evaluator's tiredness, mental state, and objectivity, makes it more effective to assign this work to a system.

Although objective responses are easy for machines to evaluate, handling subjective responses presents special difficulties because of their wide vocabulary and varying lengths. The study looks into a method for assessing subjective responses that is based on textual analysis and machine learning. We investigate methods including word mover's distance, tokenization, lemmatization, TF-IDF, Bag of Words, word2vec, similarity measurement, and cosine similarity. To assess the effectiveness of different models, the research use assessment metrics such as F1-score, Accuracy, and Recall. It also goes over other approaches used in the past to evaluate text similarity and subjective responses. The article notes that dealing with arbitrary responses, such as frequent synonyms, a broad range of durations, and unpredictable sentence sequences, has its downsides.

2 Literature Survey

The investigation of subjective response evaluation is not a new idea; over 20 years of research have explored several approaches. To tackle this problem, a variety of methods have been employed, such as the Bayes theorem, K-nearest classifier, big-data natural language processing, latent semantic analysis, and formal approaches like formal concept analysis. It has also been investigated how statistical methods, information extraction.(Wang, Ellul, & Azzopardi, 2020).Analysing criminal data stored in an organized manner from numerous sources to identify patterns and trends in crimes is one prominent use of data mining that has emerged in recent years. This method helps to automatically identify and notify about crimes, which improves the effectiveness of the processes involved in solving crimes.

The study examines the literature on various uses of data mining in crime-solving, underlines current difficulties and research gaps, and highlights the significance of selecting appropriate data mining approaches to increase the efficacy of crime detection. (Han et al., 2021)

To do this, they used data gathered by Google Research's Crowd Source team in 2019 to improve a pre-trained BERT model on our problem. According to Annamoradnejad, Fazli, and Habibi's (2020) analysis, following two training epochs, the model's Mean-Squared-Error (MSE) value was 0.046 and did not significantly reduce in the subsequent ones. The findings indicate that by fine-tuning, we may generate accurate models more rapidly and with fewer data points. Community Q&A sites such as Quora and Stack Overflow have tight guidelines that users must follow in order to preserve the content's integrity. These systems mostly require community user reports to assess material, which has serious flaws like slow violation resolution, lost time for frequent and experienced users, inadequate quality in certain reports.

Muangprathub, Kajornkasirat, and Wanichsombat's (2021) project aims to establish educational institutions that offer a range of exams annually, including competitive exams that are extra- and institutionally-based. In an effort to lessen the tension related with the exam evaluation process, online exams and tests are becoming more and more prevalent. The online assessments may consist of multiple-choice or objective questions. Nevertheless, there are solely multiple-choice or objective questions on the exams. They are creating an online system for subjective response verification based on artificial intelligence for use in all industries, including education (colleges, universities, and schools). Because it saves time and eliminates the headache of grading a mountain of papers, the suggested technique may prove to be very helpful to educators anytime they need to conduct a quick exam for revision purposes. (Bashir et al., 2021; Sakhapara et al., 2019).

Xia et al.'s (2019) project aims to establish educational institutions that administer a range of exams annually, including competitive exams that are extra- and institutionally-based. A framework to build multilanguage text corpus has been proposed by Anusha, Vasumathi, and Mittal's (2023). The COVID-19 pandemic of 2020 prompted educational systems to shift from traditional in-person learning to virtual education. (Jafar et al., 2022). Virtual classroom platforms are being used by higher education institutes (HEI) to educate in online contexts using information technology resources. (Mittal, Kaur, & Jain, 2022). Further, these digital resources are also promoting entrepreneurship skills. (Mittal, Kaur, & Gupta, 2021). There onwards online exams and tests are becoming even more common these days to ease the strain of the exam evaluation procedure. The questions on the online tests might be either objective or multiple-choice. Although AI quizzes enhances the self-regulated learning process. (Wang et al., 2023). However, the tests only have multiple-choice or objective questions. Researchers are developing an online system for ar-

tificial intelligence-based subjective answer verification in all fields, including education (schools, colleges, and universities). The proposed technique could therefore be of great use to educators whenever they need to administer a fast test for the purpose of revision, as it saves time and the hassle of grading the stack of papers.

One of the most crucial aspects of the teaching and learning process continues to be the evaluation of the responses. The need for automatic evaluation of the responses has led to the development of numerous systems in the digital age. The subjective responses often come in either short form or lengthy form. Their results of reviewing and rating the replies using the current system available for evaluation have been middling. Such scoring frameworks use data recovery techniques to compare the answers provided by students and those of references, but the results are still not ideal. (Kumari, Godbole, & Sharma, 2023)

This is due to the fact that such questions can be accurately graded by a machine. However, there are still issues with establishing an appropriate computerized grading for the ambiguous questions. Using two algorithms—latent semantic analysis and information gain for the generation of grades they built and implemented machine learning-based subjective answer grader system in this essay to deal with this problem. Converting raw text input into a numerical format that machine learning algorithms can understand is the process of feature extraction. (Bashir et al., 2021).

3 Methodologies Used

- **Data Collection:** At now, Although the suggested model needs a large corpus of subjective question replies for training and testing, there is no publicly available labelled corpus. In response, this effort, which focuses on websites and blogs with a variety of questions and replies, develops a tagged corpus of subjective answers. Data from a range of sources, including computer science and general knowledge, are gathered via web crawling.
- **Data Annotation:** More data annotation is necessary because the crawled data does not have labels. For this activity, a varied group of volunteers is chosen from our corpus of subjective question and answer data. Thirty annotators participate, comprising instructors and students from different institutes and places in Pakistan. The annotators, who are between the ages of 21 and 51, seek to accurately evaluate the subjective answers from the pupils.
- **Preprocessing Module:** Preprocessing is applied to the response and the solution after user input. Tokenization, stemming, lemmatization, stop word removal, case folding, and locating and utilizing synonyms within the text are all included in this. Interestingly, stop words that are fed to word2vec to improve semantic meaning are kept. But because they can make it more difficult to identify patterns, Before the data is input

into machine learning models such as Multinomial Naive Bayes, these stop words are eliminated.

- **Result Predicting Model:** The central component of this study is the Result Predicting Module. It works by making predictions about the outcomes of the data processing.
- **Final Score Prediction Model:** This module uses data from the machine learning module to validate the final score using learned class information. The result is deemed complete if the grade is in line with the class. The suggested score is modified according to whether it exceeds or falls short of the similarity equivalent score. Half of the values in that range are added or subtracted if the class and score do not match.

4 Architecture

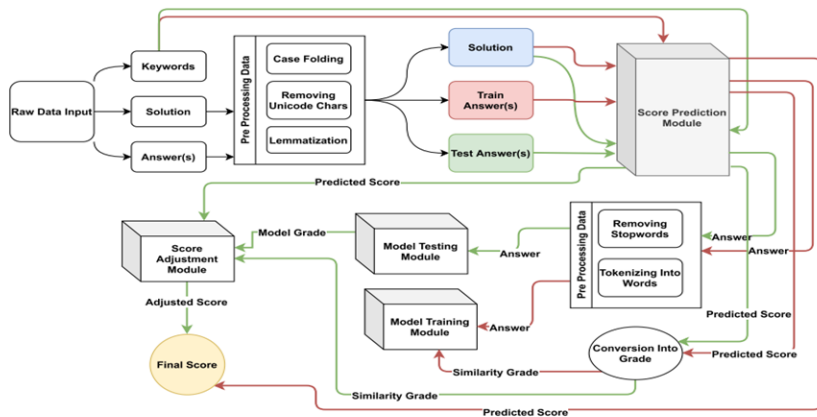


Figure 1. Architecture module for subjective answer evaluation using NLP

The architecture has been described in this section:(see figure 1)

1. **Keywords:** To answer the question in a way that is pertinent, you must use keywords. It's possible that these keywords only have the most crucial terms in lowercase. Their importance can be found in the significant influence they have on the score that the similarity evaluation module assigns.
2. **Solution:** The solution acts as a guide for mapping student responses and represents a wholly subjective reaction. All of the keywords and scenarios covered in the responses are included, each in its own line or paragraph. The answer, which is usually created

by the instructor or assessor, offers a standard by which student solutions are judged.

3. Answer: The response is a student's evaluation-subjective subjective statement. Depending on the nature of the inquiry and the student's writing style, the query may contain all or part of the keywords. requiring more semantic accuracy in processing.
4. Data Collection: Although the suggested model requires a sizable corpus of arbitrary question answers for testing and training, no publicly available labelled corpus is available at this time. In response, this effort, which focuses on websites and blogs with a variety of questions and replies, develops a tagged corpus of subjective answers. Data from a range of sources, including computer science and general knowledge, are gathered via web crawling.
5. Data Annotation: More data annotation is necessary because the data that was crawled does not have labels. For this activity, a varied group of volunteers is chosen from our corpus of subjective question and answer data. Thirty annotators participate, comprising instructors and students from different institutes and places in Pakistan. With ages ranging from 21 to 51, the annotators seek to provide accurate scores for students' subjective responses.
6. Preprocessing Module: Preprocessing is applied to the response and the solution after user input. Tokenization, stemming, lemmatization, stop word removal, case folding, and locating and utilizing synonyms within the text are all included in this. Interestingly, stop words that are fed to word2vec to improve semantic meaning are kept. However, these stop words are eliminated before the data is fed into machine learning models such as Multinomial Naive Bayes, because they may make it more challenging to spot patterns.
7. Model for Predicting Results: As Figure 3 ??illustrates, the Result Predicting Module is the main element of our investigation. It works by making predictions about the outcomes of the data processing.
8. Final Score Prediction Model: As seen in Figure 4??, this module uses learnt class information to validate the final score using data from the machine learning module. The result is deemed complete if the grade is in line with the class. The suggested score is modified according to whether it exceeds or falls short of the similarity equivalent score. Half of the values in that range are added or subtracted if the class and score do not match.

The corrected score post-model recommendation, which takes into account errors from the Score Prediction and Machine Learning Module, is accepted as final in cases where

the machine learning model has received significant training; in the event that the model has not received enough training, the score is assumed to be correct.

5 Flow Chart

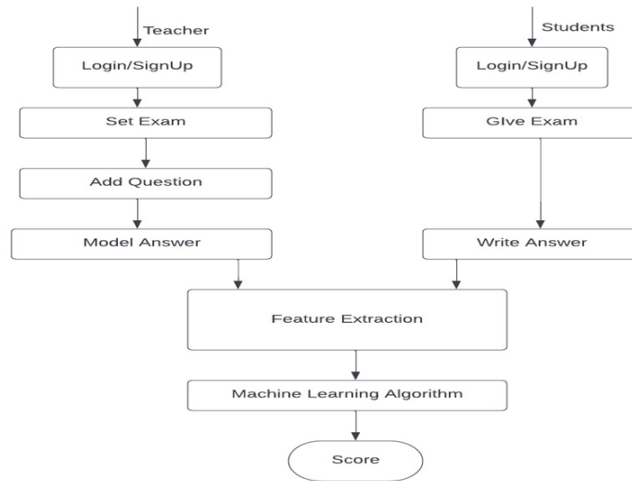


Figure 2. Workflow of Subjective answer evaluation using NLP

Our system is divided into two portions, one for teachers and the other for students. The teacher can login page if they are not already enrolled. In order to configure the exam paper in the teachers’ section, they must first establish a test by entering the test name, date, and time. After creating the test, the teacher must add the questions, along with the appropriate marks and sample answers. To take the exam in the student part, the student must enroll in the class. The student’s answers will be compared to the model answers given by the paper setter after they have taken the test. These responses will be evaluated according to their cosine similarity to the model response, answer length, keyword check, grammar check, context & semantic similarity, and grammar check.(see figure 2).

The following are the system’s key components:

1. Feature Extraction: Involves converting raw text data into a structured format, enabling machines to understand, analyze, and make sense of the textual information. Such as text classification, sentiment analysis, and information retrieval, feature extraction plays a pivotal role in representing the inherent characteristics of the text in a way that machine learning algorithms can effectively work with.

2. Machine Learning Algorithm: Machine learning algorithms are employed to automatically evaluate and classify subjective answers written in natural language. Like Naive Bayes, Decision trees, these algorithms analyze various linguistic and contextual features of the answers to assign appropriate scores or labels, thereby streamlining the assessment process.
3. Final Score Prediction: The final score prediction emerges as a numerical output that represents the model's estimation of the quality, depth, and correctness of the response. This prediction is based on the insights gained from analyzing the text's content, grammar, coherence, vocabulary, and any other relevant features. The prediction process essentially automates what would traditionally be a manual and subjective task, providing educators with an objective and consistent assessment tool.

6 Result

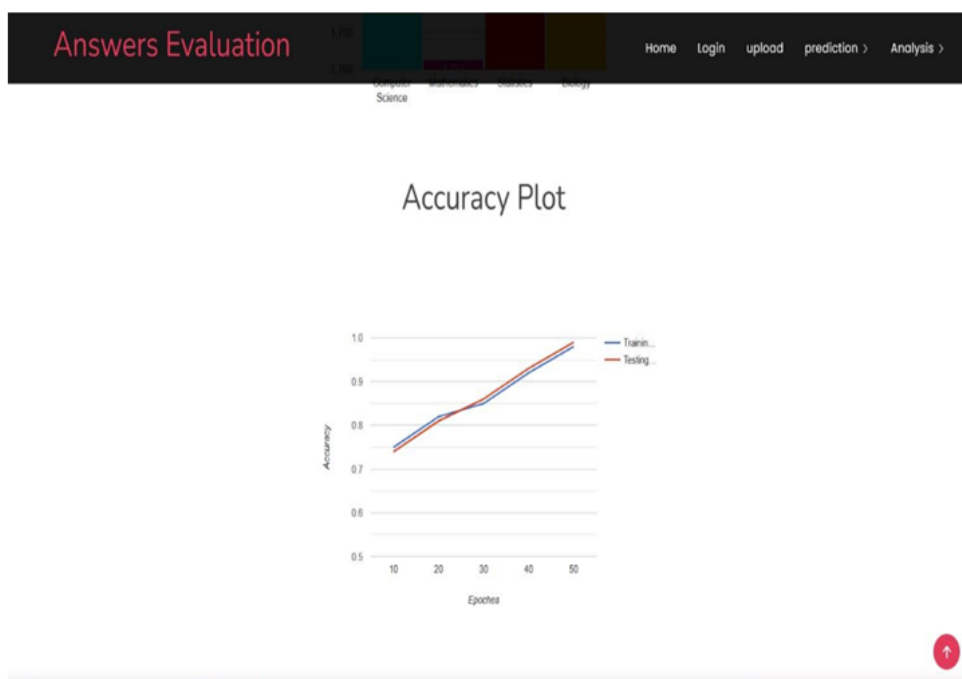


Figure 3. Working of Subjective Answer Evaluation Using NLP Accuracy Plot

There are various phases involved in applying NLP and deep learning to evaluate subjective responses explained below:

- Image-to-Text Conversion: Initially, the system converts answer scripts based on images into editable text format by means of Optical Character Recognition (OCR).
- NLP-Based Word Embeddings: Subsequently, word embedding vectors are created from both the answer key texts and the answer script using sentence transformers, a method of Natural Language Processing. The similarity measurements between these vectors are then determined by matching them using methods such as fuzzy search, Spearman's rank-order correlation, and BERT encoding.
- Performance Evaluation: F1-score, precision, recall, and accuracy are used to gauge how well the suggested model performs.(see figure 3).
- By automating the evaluation process, this method seeks to increase its accuracy and efficiency. Evaluators stand to gain from less manual labour and quicker results.

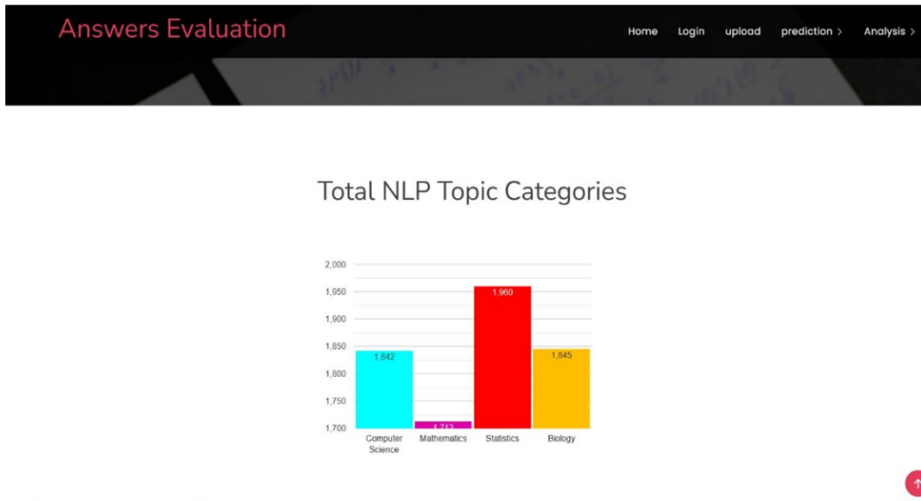


Figure 4. Working of Subjective Answer Evaluation Using NLP Total NLP topic categories

Therefore, letting a system handle this time-consuming and occasionally crucial duty of analyzing subjective responses is significantly more time and resource-efficient in various domains and categories.(see figure 4). The evaluation of objective responses by machines is rather easy and useful. One-word answers to questions can be supplied to a programmer that can quickly map students' responses. However, handling subjective answers is much trickier. They range widely in length and have a sizable vocabulary. We look

into a method that makes advantage of machine learning and natural language processing to assess subjective responses. Tokenization, lemmatization, text representation, Bag of Words, word2vec, similarity measurement, cosine similarity, and word mover's distance are some of the methods used in our research to handle natural language.

7 Conclusion

The "Subjective Answer Evaluation using NLP" project is a noteworthy accomplishment in the realm of educational assessment since it blends the power of Natural Language Processing (NLP) with the intricate subtleties of subjective answer evaluation. By building a sophisticated automated system, this project addresses the consistency, efficiency, and objectivity issues that have long dogged the assessment process. The project gives teachers a tool that uses innovative scoring algorithms and state-of-the-art NLP models to ensure accurate and comprehensive evaluation of subjective responses while also relieving the burden of human grading. The system is a flexible solution that can be tailored for various courses, levels, and languages due to its ability to adjust to various educational scenarios. Its applicability in both traditional classroom settings and virtual learning environments guarantees its relevance in the evolving field of education.

This undertaking has ramifications beyond what educators will feel in the near future. It improves students' learning experiences by providing them with timely and constructive feedback that points them in the direction of academic growth. Because NLP technology is seamlessly integrated into the evaluation process, students are better able to discover their areas of strength and development. This facilitates a culture of lifelong learning and self-improvement. It's critical that we continue to establish a balance between the benefits of technology and the insights that come from human judgment. Though it should be used as a tool to empower teachers and enhance the learning process rather than taking the place of their experience, NLP-based evaluation has a lot of potential.

Finally, NLP-driven subjective answer evaluation provides opportunities for improved consistency, speed, and perceptive feedback within the framework of assessment in education. Through a responsible and comprehensive adoption of this technology, we may build a future in which evaluation is an AI and human intelligence combined effort.

References

Annamoradnejad, I., Fazli, M., & Habibi, J. (2020). Predicting Subjective Features from Questions on QA Websites using BERT. 2020 6th International Conference on Web Research, ICWR 2020, 240–244. <https://doi.org/10.1109/ICWR49608.2020.9122318>

- Anusha, K., Vasumathi, D., & Mittal, P. (2023). A Framework to Build and Clean Multi-language Text Corpus for Emotion Detection using Machine Learning. *Journal of Theoretical and Applied Information Technology*, 101(3), 1344–1350.
- Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective Answers Evaluation Using Machine Learning and Natural Language Processing. *IEEE Access*, 9, 158972–158983. <https://doi.org/10.1109/ACCESS.2021.3130902>
- Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., & Gao, C. (2021). A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, 33(5). <https://doi.org/10.1002/cpe.5971>
- Jafar, A., Dollah, R., Dambul, R., Mittal, P., Ahmad, S. A., Sakke, N., Mapa, M. T., Joko, E. P., Eboy, O. V., Jamru, L. R., & Wahab, A. A. (2022). Virtual Learning during COVID-19: Exploring Challenges and Identifying Highly Vulnerable Groups Based on Location. *International Journal of Environmental Research and Public Health*, 19(17). <https://doi.org/10.3390/ijerph191711108>
- Kumari, V., Godbole, P., & Sharma, Y. (2023). Automatic Subjective Answer Evaluation. *International Conference on Pattern Recognition Applications and Methods*, 1, 289–295. <https://doi.org/10.5220/0011656000003411>
- Mittal, P., Kaur, A., & Gupta, P. K. (2021). THE MEDIATING ROLE of BIG DATA to INFLUENCE PRACTITIONERS to USE FORENSIC ACCOUNTING for FRAUD DETECTION. *European Journal of Business Science and Technology*, 7(1), 47–58. <https://doi.org/10.11118/ejobsat.2021.009>
- Mittal, P., Kaur, A., & Jain, R. (2022). Online Learning for Enhancing Employability Skills in Higher Education Students: The Mediating Role Of Learning Analytics. *TEM Journal*, 11(4), 1469–1476. <https://doi.org/10.18421/TEM114-06>
- Muangprathub, J., Kajornkasirat, S., & Wanichsombat, A. (2021). Document plagiarism detection using a new concept similarity in formal concept analysis. *Journal of Applied Mathematics*, 1–10. <https://doi.org/10.1155/2021/6662984>
- Sakhapara, A., Pawade, D., Chaudhari, B., Gada, R., Mishra, A., & Bhanushali, S. (2019). Subjective answer grader system based on machine learning. *Advances in Intelligent Systems and Computing*, 898, 347–355. https://doi.org/10.1007/978-981-13-3393-4_36
- Wang, X., Wrede, S. E., van Rijn, L., & Wöhrle, J. (2023). Ai-Based Quiz System for Personalised Learning. *ICERI2023 Proceedings*, 1, 5025–5034. <https://doi.org/10.21125/iceri.2023.1257>
- Wang, X., Ellul, J., & Azzopardi, G. (2020). Elderly Fall Detection Systems: A Literature Survey. *Frontiers in Robotics and AI*, 7. <https://doi.org/10.3389/frobt.2020.00071>

Xia, C., He, T., Li, W., Qin, Z., & Zou, Z. (2019). Similarity Analysis of Law Documents Based on Word2vec. Proceedings - Companion of the 19th IEEE International Conference on Software Quality, Reliability and Security, QRS-C 2019, 354–357. <https://doi.org/10.1109/QRS-C.2019.00072>