A PLATFORM VIEWW OF AUTIMATIC SHORT ANSWER SCORING SYSTEM

Yunita Sari
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Guntur Budi Herwanto
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Mardhani Riasetiawan
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Bambang Nurcahyo Prastowo
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Isna Alfi Bustoni
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Indra Hidayatullah
Universitas Negeri Yogyakarta

Safari
Pusat Penelitian Pendidikan, Kemdikbud Republik Indonesia

Edgar Anarossi
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

I Nyoman Prayana Trisna
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Kurniawan Eka Nugraha
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

Harista Sriwijayuni
Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada

ABSTRACT

In this paper we present a platform for automatic short answer scoring in Bahasa Indonesia. This platform was built based on Machine Learning model which learns how human evaluator assigning score. This platform is not merely an automatic scoring machine, but it also enables users to re-train the machine until the optimal performance is obtained. The platform provides some features that can be customized to the user needs starting from the pre-processing to the choice of Machine Learning models. We named the platform UKARA, a javanesse word for a sentence. UKARA Automatic Short Answer Scoring (ASAS) has been implemented with FastText, Support Vector Machine, and AdaBoost to learn from the human evaluation. Above all, UKARA ASAS comes with a user-friendly interface which allows non-technical users to use it without the hassle.

KEYWORDS

Short Answer. Bahasa Indonesia, Machine Learning, UKARA, Scoring System
1. INTRODUCTION

The development of the world of technology currently touches in various aspects of life, including the world of education. Over the past few years, technology of all types has played a major role in efforts to improve the quality of education, both in terms of teaching and assessment. In terms of teaching, e-Learning technology and Learning Management System are mushrooming in various parts of the world. In terms of assessment the most famous and widely used example is the scantron machine that is relied on by teachers in the assessment of multiple-choice questions. In Indonesia, this machine has become a mainstay in the National Examination.

Since 2014, national-scale exams have been tried in the form of Computer Based Testing in Singapore Indonesia Middle School and Kuala Lumpur Indonesia Middle School. The results of the implementation of UNBK (CBT) showed a positive response. Therefore, the UNBK system began to be implemented gradually in schools in Indonesia. Through the UNBK system, exam questions can be made not only in the form of multiple choice but also in the form of short answers and essays.

However, the implementation of UNBK with the concept of short answer and essay presents new challenges for the world of education. The assessment process must be free from fraud, subjectivity as much as possible, and can be done quickly and accurately. Therefore, we need a system that automatically evaluates the short answer answers and student essays in the UNBK exam.

We propose UKARA as an Automatic Short Answer Scoring platform that can help the process of evaluating student answers in UNBK (Herwanto et al, 2018). UKARA is a machine learning based platform that currently uses FastText, Support Vector Machine (SVM), and AdaBoost to "learn" to understand patterns from human evaluation. UKARA also features a model retrain to achieve the highest accuracy. In this paper we present information on the development of the UKARA platform for both evaluators and graders.

2. RELATED WORK

Research on the automatic essay grading system has been developed since 1966 by Ellis Page under the name Project Essay Grader (PEG) which was later criticized for being too focused on the surface structure and ignoring the semantic aspects of the essay (Islam & Hoque, 2013). An automatic essay grading system continues to develop until now, one of which is an automatic essay grading system named ABESS (Islam & Hoque, 2013), applying the Generalized Latent Semantic Analysis (GLSA) method with the Singular Value Decomposition (SVD) algorithm and Cosine Similarity.

The GLSA method was chosen because it is a form of IR-technique development from the Latent Semantic Analysis (LSA) method. The data used in this system is in the form of essays in Bangla that have previously been through the training stage. The system accuracy rate of 98% works in the form of plain text essays.

On the UKARA platform the FastText, SVM, and AdaBoost classifier is used. FastText is a classifier with skipgram model approach. In this beta version of UKARA, UKARA was implemented at UNBK for Indonesian subjects. Indonesian, as one of the subjects tested in UNBK, has at least 127 thousand entries (Saputra, 2018).

The concept of FastText itself is to store words in the form of the sum of n-gram vector representations for each word (Bojanowski, 2016). With this concept, data processing becomes faster and has good performance for data with languages that have a lot of vocabulary. Therefore, FastText was chosen to be implemented in UKARA.

The SVM method is one of the most commonly used methods in text classification. Several studies have implemented the SVM method and produced good performance. Research using SVM in scoring essays for GRE Analytical Writing through Education Testing Service (ETS) results in
60% accuracy for one-to-one cases, but by allowing one-point error to increase accuracy up to 92% (Suresh, 2018).

3. UKARA PLATFORM

UKARA is a system that automatically classify freely written text into several set of grades (Herwanto, 2017). The system is built for evaluator or grader. We build UKARA as a web-based application so it can be used without installation. In this paper, we want to present a complete overview of our system, including the functionality of each module and the interaction between the modules. We describe the interaction between the modules in Figure 1.

1. Create question and answer set
   This is the first step of modeling automatic short answer scoring (ASAS). The ASAS model is independent for each question. The evaluator needs to populate data such as question name, the question itself, rubrics, and answer set. After that, there will be an upload window that will generate excel files to be filled as answer set. We provide an excel files, because this is the most common export format for learning management system such as moodle, canvas, etc. The evaluator can move the data through the excel provided by UKARA ASAS, and then import it to the system.

2. Similarity Checking and Manual Labeling
   This module provides evaluator a hint about the similarity between the answer and the answer key or rubrics. The aim of this module is to help human label the answer data based on the similarity with example answers. Currently we are using Bilingual Evaluation Understudy (BLEU) score to give the similarity score, but we are looking for more robust alternatives such as cosine or jaccard similarity. The manual labeling doesn’t have to be conducted on all the answer set. Instead, our goal is to make the evaluator to be only give 10% of manual labeling from all the answer set. While the score of the other answer will be inferred from the AES model.

3. Preprocessing
   Prior to the automatic scoring is preprocessing step. Pre-processing doesn’t always produce the best result, so we make this as on option for the evaluator. In UKARA ASAS, we propose to preprocess method which is stopword removal and stemming. Stopword is a common word that frequently appear in the text but doesn’t provide a meaningful context to the text. These words sometimes different between cases, so the platform provides a free text input for the user to fill additional stopwords. Whereas stemming is the process for finding the root of the word by removing the prefix and suffix while finding the matching word in the dictionary. This process doesn’t always produce the best result in the final performance, so we make this as on option that can be chosen by the evaluator.

4. Model Training
   There are two main steps on this module which is feature representation and the classification process. For the feature representation we rely on the classic term frequency and inverse document frequency method. This method proves to perform best on the different kind of situation, and language independent, which we want on this platform. This method basically gives a larger weight for the words that appear a lot in a document but rarely appear in other documents, and vice versa. The text will be converted into vector based on this intuition. After that, the vector will become the input for the classifier method. We propose 4 methods, which is Naive Bayes, Support Vector Machine, AdaBoost, and also RandomForest. We also give this as an option for the evaluator to choose, because the performance can be different between cases. The output of the model training is the performance between 0-100. The greater the number, the better system can predict the answer set to the real score. In our experience, the question with the short answer type gives the performance about 98%, but in the question with the longer and analytic answer, the performance only reach 70-80%.
5. Inference
   Once the evaluator find the best performance, they can start an inference. This is where the automatic scoring conducted. As we said in the manual labeling, the answer without score will be automatically scored from the model that has been produced in the previous steps. Evaluator can add more answer set, on this steps, and let the system do the automatic scoring.

6. Re-training
   The performance of the ASAS is rarely reach 100%. This is because there are some false positives (fp) and false negative (fn). False positive is when the wrong answer predicted as a right answer, and false negative is when the right answer predicted as a wrong answer. Based on our findings, these mistakes can be happening because of the error in manual labeling. This platform will display the fp and fn that can be evaluated by the evaluator, and later can be corrected to be re-train and hopefully give the better performance. The evaluator can also add more manual label to the system, that can increase the performance of the current model.

![UKARA Platform Modules Diagram](image_url)
4. CASE STUDY

UKARA has been tested using several types of datasets which were developed from the student’s answer of PISA (Programme for International Student Assessment) questions. In this stage, we are focusing on the type of questions which fall on the global competence category. The purpose of the questions in this category is to examine the student’s perspectives on facing local, global and intercultural issues, their openness towards people from other cultural backgrounds and global mindedness. One of the characteristics of global competence questions is it challenge the students to express their critical thinking towards the stimulus and question given. Table 1 show a set of global competence question which consists of stimulus, question and coding guidelines. All of them have been translated into Bahasa Indonesia.

Table 1. An Example of Global Competence question

| STIMULUS | Sebuah toko baju berkonsep self-service menawarkan promosi dua buah baju bertema tahun baru seharga Rp50.000,00. Sebelum baju bertema tahun baru dibagikan kepada pembeli, sebuah layar akan menampilkan tampilan gambar yang menampilkan kondisi kerja di dalam sebuah pabrik konveksi/pembuatan baju. Kemudian pembeli diberi program pilihan untuk menyelesaikan pembeliannya atau menyumbangkan Rp50.000,00 untuk dijadikan donasi pembagian baju musim dingin di suatu daerah yang membutuhkan. Delapan dari sepuluh pembeli memilih untuk memberikan donasi. |
| QUESTION | Menurut anda mengapa banyak dari pembeli yang memilih berdonasi? |
| CODING GUIDELINES | Label 1: Siswa memberikan jawaban yang menjelaskan alasan untuk melakukan donasi yang merujuk pada kesadaran akan kondisi kerja di industri pakaian atau bagaimana tindakan konsumen mempengaruhi orang lain ATAU siswa menjelaskan alasan untuk melakukan donasi yang berfokus pada emosi atau motivasi dari donor saja. Label 0: Siswa memberikan jawaban yang salah, tidak jelas, tidak cukup atau tidak relevan. |

The stimulus explained a situation when someone was about to buy a t-shirt. Before the buyer starts the purchase, a video showing a situation in the clothing manufacturer will be played. The buyer then will be given a choice to finish the purchase or to donate the money to charity. Eight of ten buyers decided to donate their money. The question is why more buyers choose to donate their money? Students will be assigned a score of 1 if their answer related to the awareness of the working conditions in the labour intensive industry or if their answer focus on the motivation of the donation. A score of 0 will be given if the students give wrong, unclear, or irrelevant answer.

In order to answer this type of question, the students are required to have specific cognitive and socio-emotional skills including reasoning with information, communication in intercultural contexts, perspective-taking, conflict resolution and adaptability. In addition to that, students need to answer in a concise manner. This type of question is a challenge for UKARA, as the system will be given a variety of answers in relatively wide spectrum. As explained in the previous section, UKARA works by learning from a set of training data (scored student’s answers). There is always a risk that the distribution of the answers in the training data is different from the answer of un-scored answers, which may lead to poor performance of the system. Table 2 shows the sample of student answers which have been scored manually by the human grader.
Table 2. Sample of student’s answers

| id | Sample answer                                                                 | Score |
|----|-------------------------------------------------------------------------------|-------|
| 1  | Karena orang berpikir bahwa jika disumbangkan akan membuat produksi pakaian menjadi lebih beretika | 1     |
| 2  | karna orang lebih suka menyumbang                                             | 1     |
| 3  | karena dia kasihan                                                             | 1     |
| 4  | karena menyumbang itu harus di lakukan karena wajib dan akan mendapatkan yang lebih untuk orang yang menyumbang | 1     |
| 5  | dikarenakan kondisi pabrik pembuat pakaian yang mengejutkan                    | 1     |
| 6  | karena kondisi di sebuah pabrik pakaian yang sangat mengejutkan                | 0     |
| 7  | karena pakaian yang tidak dipakai lagi lebih baik disumbangkan dari pada dibuang di tempat sampah. | 0     |
| 8  | karana harga nya terjangkau dan pas.                                         | 0     |
| 9  | karena yang tak terpakai lebih beretika                                        | 0     |
| 10 | karna lebih baik menyumbang daripada yang disumbang,menyumbang kita dapat pahala atau rizki dari allah | 0     |

The table shows a variety of answers which reflects the cognitive and socio-emotional skills of the students. Some answers are heavily influenced by the student’s perspective of giving donation (i.e religious perspective (id: 10)). By looking to the sample answer, the challenge might start from the process of developing a reliable training data. Human graders need to carefully perceive the conveyed idea of the students which can be ambiguous. As an example, the sample answers of id: 4 and 10 are fundamentally has similar ideas. However, the human grader assigned different scores to both answers. A similar thing is found in the sample answer of id:4 and 5. Another interesting finding can also be seen in the sample answers of id: 3 and 6. The student that has answer of id:3, has a simple perspective that donation should be given because the buyers feel pity to the workers in the clothing manufacturer. Compare to the answer of the id:6, the student may have more critical thinking towards the condition in the manufacturer. However, since it does not mention any keywords related to donation, the human grader assigned 0 score.

By looking to sample answer in Table X, we are aware that another challenge is typographical errors. As an example, the word karena is written in different variations: karna, karana, Karena, KARENA. This problem might be trivial, however due to the cultural influence (i.e mother language), we may have more number of variations which should be addressed properly. We also found that it is much easier to comprehend the original version of the problem that was written in English. This is another challenge to The Center for Educational Assessment (Puspendik) to create their own version of global competence problems.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the platform view of UKARA automatic short answer scoring (ASAS). We described the details of all the modules available in UKARA. In addition, we discussed a case study which shows the challenges remains that need to be addressed. In the future, we are planning to implement word embedding representation which has the ability to capture semantic information. This representation is expected to improve the performance of our system especially for high order thinking problems and longer answers. In addition to that, we are working
on the typographical corrections and building a faster stemming algorithm. Finally, we aim to build a more friendly user interface, especially for non-technical users.

ACKNOWLEDGEMENT

The research has supported by PUSPENDIK KEMDIKBUD Republic of Indonesia.

REFERENCES

Herwanto, GB. Sari, Yunita. Prastowo, BN, Bustoni, IA, Hidayatulloh, Indra. (2018). UKARA: A Fast and Simple Automatic Short Answer Scoring System for Bahasa Indonesia. Proceeding Book of 1st 2018 International Conference on Educational Assessment and Policy, Vol 1.

Islam, M. M., & Hoque, A. S. M. L. (2013). Automated Bangla essay scoring system: ABESS. 2013 International Conference on Informatics, Electronics and Vision (ICIEV), 1–5. https://doi.org/10.1109/ICIEV.2013.6572694

Kemendikbud, https://unbk.kemdikbud.go.id/tentang

Bojanowski, Piotr. Grave, Edouard. Joulin, Armand. Mikolov, Tomas. (2016). Enriching Word Vectors with Subword Information. 2016 Transactions of the Association for Computational Linguistics,

Saputra, Reza Deni. (2018). Badan Bahasa Mutakhirkan KBBI Daring Pertama Kali. https://www.liputan6.com/news/read/3440643/badan-bahasa-mutakhirkan-kbbi-daring-pertama-kali

Suresh, Abhishek; Jha, Manuj. (2018). Automated Essay Grading using Natural Language Processing and Support Vector Machine. International Journal of Computing and Technology, Volume 5, Issue 2, February 2018.