INTISARI

Nilai adalah indeks yang digunakan oleh para pengajar untuk mengetahui tingkat pemahaman peserta didik terhadap materi yang telah diberikan. Salah satu metode pengujian yang digunakan adalah soal dengan jawaban berupa esai singkat, karena lebih mampu menggali tingkat pemahaman siswa secara mendalam. Metode esai singkat menimbulkan masalah berupa peningkatan kompleksitas dan waktu penilaian. Jawaban esai singkat yang subjektif mengakibatkan kriteria penilaian sulit untuk direpresentasikan secara kuantitatif dan akurat. Akibatnya, hasil penilaian pun dapat berbeda, bahkan antar penilai manusia. Pendekatan yang dikembangkan untuk menyelesaikan masalah koreksi jawaban esai adalah sistem penilaian esai otomatis (*automated essay scoring system*), yang dapat mendeteksi tiap kata dalam esai lalu mencocokkan kemiripannya dengan kunci jawaban yang telah disediakan oleh pembuat soal menggunakan teknologi kecerdasan buatan. Akan tetapi, sistem ini kebanyakan dikembangkan untuk menilai esai singkat dalam bahasa asing, seperti bahasa Inggris.

Solusi yang diajukan untuk permasalahan ini adalah pengembangan sistem penilaian esai singkat menggunakan metode *Quadratic Weighted Kappa* (QWK), yang dipilih karena dapat memvalidasi penilaian dari dua pihak penilai sehingga hasil penilaian menjadi lebih andal dan akurat. Sistem penilaian juga menambahkan jawaban mahasiswa dengan nilai similaritas tertinggi ke kunci jawaban untuk menambah cakupannya. Sistem dikembangkan untuk *platform desktop* dengan sistem operasi Windows menggunakan bahasa pemrograman Python. Proses pengembangan proyek *capstone* dilakukan menggunakan model *waterfall*.

Sistem penilaian esai otomatis yang dikembangkan telah diuji pada tiga jenis *dataset*, yang masing-masing terdiri atas 138, 215, dan 243 data. Hasil pengujian menunjukkan bahwa penambahan jawaban mahasiswa ke kunci jawaban dapat meningkatkan hasil similaritas sehingga juga meningkatkan hasil validasi QWK hingga sebesar 15,96%. Sistem ini masih memiliki beberapa kekurangan, salah satunya belum bisa menilai similaritas dengan akurat. Nilai QWK tertinggi pada pengujian hanya mencapai 0,2785 atau 27,85%, yang berada pada kategori “cukup ada persetujuan”. Hal ini disebabkan oleh dua faktor utama, yaitu banyaknya data dalam *dataset* dan algoritma pengecekan similaritas yang masih naif. Semakin banyak jumlah datanya maka nilai QWK semakin rendah. Oleh karena itu, perlu implementasi algoritma pengecekan similaritas yang dapat menilai kemiripan jawaban mahasiswa dan kunci jawaban dengan lebih akurat serta dapat menilai data berjumlah banyak.

Kata Kunci: Bahasa Indonesia, esai pendek, kecerdasan buatan, sistem penilaian otomatis, *Quadratic Weigthed Kappa (QWK)*.

ABSTRACT

Grades are an index used by educators to determine students' level of understanding of the given learning material. One of the testing methods used is questions with answers in the form of short essays because it is more capable of probing students' level of understanding in depth. However, the short essay method poses problems in the form of increased complexity and assessment time. The subjective nature of short essay answers makes it difficult to represent the assessment criteria quantitatively and accurately. As a result, assessment results can differ, even between human graders. An approach developed to solve the essay answer assessment problem is an automated essay scoring system, which can detect each word in the essay and match its similarity with the answer key provided by the question maker using artificial intelligence technology. However, these systems are mainly developed to score short essays in foreign languages, such as English.

The proposed solution to this problem is developing a short essay scoring system using the Quadratic Weighted Kappa (QWK) method, which can validate the assessment from two sides of assessors so that the assessment results become more reliable and accurate. The scoring system also adds student answers with the highest similarity scores to the answer key to increase its coverage. The system is developed for desktop-platform with Windows operating system using Python programming language. The capstone project is implemented using the waterfall model.

The developed automated essay scoring system has been tested on three datasets, consisting of 138, 215, and 243 data, respectively. The test results show that adding student answers to the answer key can improve the similarity results, which consequently increases the QWK validation results by up to 15.96%. However, this system still has some shortcomings, one of which is that it cannot accurately assess similarity. The highest QWK score in the test only reached 0.2785 or 27.85%, which is in the " fair agreement" category. This shortcoming is due to two main factors: the amount of data in the dataset and the naïve similarity checking algorithm. The tests show that the larger the data, the lower the QWK value. Therefore, it is necessary to implement a similarity-checking algorithm that can assess the similarity of student answers and answer keys more accurately and can assess large amounts of data.

Keywords: Indonesian language, short essays, artificial intelligence, automated scoring system, Quadratic Weighted Kappa (QWK).