

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Sebagai *lingua franca* dunia, penguasaan bahasa Inggris tidak hanya digunakan sebagai persyaratan akademik dan juga tidak terbatas pada pengetahuan terhadap bahasa tetapi juga sebagai alat komunikasi *global* (Shobikah, 2017). Hingga saat ini bahasa Inggris dianggap sebagai bahasa global yang berperan dalam interaksi dan komunikasi global. Sehingga, penguasaan bahasa Inggris merupakan salah satu modal dasar untuk dapat berkompetisi dalam menghadapi era global (Handayani, 2016).

Salah satu bentuk evaluasi pembelajaran kemampuan berbahasa Inggris adalah dengan mengikuti TOEFL atau *Test of English as a Foreign Language* yang menurut (Alderson & Hamp-lyons, 1996) TOEFL merupakan tes yang paling terkenal di bidang ELT (*English Language Teaching*). Mengutip dari halaman web *Educational Testing Service*, nilai tes TOEFL digunakan sebagai syarat administrasi penerimaan oleh lebih dari 11000 universitas dan institusi akademik di 150 negara.

Kemudian, masalah muncul ketika pertanyaan yang diajukan pada TOEFL tidak dibagikan bebas kepada peserta tes, sehingga untuk mempersiapkan ujian peserta berlatih dengan mengandalkan pertanyaan yang ada pada buku persiapan TOEFL yang jumlahnya terbatas (Susanti et al., 2015). Selain itu, soal TOEFL yang telah beredar dianggap telah usang. Topik yang terdapat pada soal-soal TOEFL merupakan topik lama yang selalu berulang ditampilkan. Atas dasar tersebut, ketersediaan sumber belajar berupa kumpulan soal yang tidak terbatas bagi peserta tes menjadi penting, kemudian alangkah lebih baik jika topik yang diangkat pada soal TOEFL disesuaikan dengan kondisi saat ini.

Pada TOEFL soal-soal yang diujikan dibagi menjadi empat bagian yaitu, *listening comprehension*, *structure and written expression*, dan *reading comprehension* (Huda, 2017). Pada bagian *reading comprehension*, peserta ujian diuji untuk memahami berbagai jenis bacaan ilmiah yang disediakan dalam bahasa Inggris. Pertanyaan yang diajukan adalah mengenai topik dari bacaan, ide utama,

isi bacaan, arti suatu kata pada bacaan dan informasi rinci yang ada. Terdapat setidaknya sepuluh tipe soal pada bagian tersebut, yaitu *vocabulary question*, *inference question*, *reference question*, *purpose question*, *detail (factual information) question*, *negative factual information question*, *essential information (sentence simplification question)*, *sentence insertion question*, *complete the summary*, dan *complete the table question*.

Vocabulary question merupakan tipe soal pilihan ganda, dimana soal pilihan ganda atau *multiple choice question* (MCQ) adalah bentuk penilaian yang sangat populer. Pada tipe soal ini responden diminta untuk memilih jawaban terbaik dari serangkaian pilihan. Soal pilihan ganda terdiri dari tiga elemen: stem, kata target, dan pengecoh. Stem (juga dikenal sebagai item) adalah kalimat dari mana pertanyaan itu dibentuk, kata target (juga disebut sebagai kunci) adalah jawaban yang benar dari pertanyaan dan pengecoh atau pengecoh adalah himpunan jawaban yang salah (Bhatia et al., 2013). Pada *vocabulary question* TOEFL diberikan sebuah teks dibaca dan dianalisis oleh peserta. Pada soal ini diberikan satu kata yang ada pada teks bacaan yang perlu dicari persamaan katanya dari 4 opsi yang diberikan, opsi tersebut biasanya setara atau sedikit memiliki makna yang sama sehingga dapat mengecoh peserta tes. Pada soal ini, kemampuan pemahaman teks dan pengetahuan *vocabulary* peserta diuji.

Pembuatan soal pilihan ganda secara manual mahal dan memakan waktu, penggunaan teknologi komputer tentunya sangat mengurangi waktu yang dihabiskan oleh pembuat soal dalam membuat soal ujian (Aldabe et al., 2006). Para peneliti telah tertarik pada pembuatan soal pilihan ganda otomatis sejak akhir 90-an, penelitian untuk membuat soal pilihan ganda otomatis masih menjadi bidang yang aktif hingga saat ini (Ch & Saha, 2020) Salah satu keilmuan dan teknologi yang dimanfaatkan dan dapat membantu adalah *Natural Language Processing* dan *Machine Learning*.

Machine Learning merupakan salah satu bidang kajian kecerdasan buatan yang mempelajari tentang pembelajaran (*learning*) pada data dan pengalaman (Willi Richert, 2013) *Natural Language Processing* (NLP) atau dalam Bahasa Indonesia disebut sebagai Pemrosesan Bahasa Alami merupakan salah satu cabang

dari kecerdasan buatan (AI) yang dihubungkan dengan bidang komputasi linguistik (Pustejovsky & Stubbs, 2013).

Penelitian yang melibatkan *Natural Language Processing* dan *Machine Learning* untuk membangkitkan soal pilihan ganda otomatis telah dilakukan oleh beberapa peneliti sebelumnya, seperti yang dilakukan (Hoshino & Nakagawa, 2005) berjudul *A real-time multiple-choice question generation for language testing – a preliminary study*– yang membuat tipe soal *fill-the-blank* pilihan ganda dengan membentuk 7 fitur dari teks bacaan, (Brown et al., 2005) pada tulisannya *Automatic Question Generation for Vocabulary Assessment* membuat 6 jenis soal yang menguji kosa kata berupa *fill the blank* dari wordnet menunjukkan pertanyaan yang dibuat otomatis memberikan ragam kosa kata yang baik dibandingkan dengan pertanyaan yang ditulis oleh manusia, lalu (Susanti et al., 2015) *Automatic Generation of English Vocabulary Tests* memfokuskan untuk membuat tipe soal *vocabulary* yang mendapatkan hasil evaluasi dari guru bahasa Inggris yang menebak 45% soal yang dibangkitkan oleh komputer adalah soal buatan manusia, lalu (Ch & Saha, 2020) *Automatic Multiple Choice Question Generation From Text: A Survey* menyimpulkan bahwa terdapat enam fase untuk membangun *automatic multiple choice question* setelah melakukan *review* terhadap penelitian yang diterbitkan dibidang ini sejak tahun 90an.

Pada penelitian ini, penulis mengusulkan sebuah sistem penghasil soal *vocabulary* yang ada pada *section reading comprehension* dalam TOEFL menggunakan algoritma *Learning Vector Quantization* (LVQ). Penggunaan algoritma LVQ bukanlah tanpa alasan, melainkan karena LVQ cocok untuk menangani masalah *statistical* dan *pattern recognition* atau klasifikasi baik itu klasifikasi *binary* dan *multiclass* klasifikasi (Kohonen, 2001). Algoritma LVQ juga memiliki keunggulan dibandingkan algoritma *instance based learning* seperti *K-Nearest Neighbor* yang digunakan pada penelitian yang dilakukan oleh (Hoshino & Nakagawa, 2005) dan (Riza et al., 2019), dimana pada LVQ perhitungan jarak untuk klasifikasi data baru hanya dilakukan terhadap *codebook* vektor yang telah dibuat, sehingga tidak perlu membandingkan dengan keseluruhan data latih.

Untuk melakukan *question generation* dari sebuah situs media berita bahasa Inggris, sebuah teks narasi berita bahasa dipisahkan menjadi data kalimat-kalimat

kemudian kalimat tersebut dilakukan *preprocessing* untuk menghilangkan karakter yang tidak dibutuhkan, lalu data kalimat dijadikan data *instance* kata yang kemudian setiap kata diekstraksi 9 fiturnya. Kemudian algoritma LVQ akan mempelajari data latih berupa data *historical* soal *vocabulary* TOEFL, sehingga terbentuklah bobot *codebook* vektor yang merupakan representasi data latih dan algoritma LVQ juga digunakan untuk penentuan kata target yang akan dijadikan soal. Kemudian dari kata target dicari jawabannya dengan bantuan Synset wordnet, lalu ditentukanlah pengecoh berdasarkan *heuristic* yang telah ditentukan.

1.2 Rumusan Masalah

Sesuai latar belakang masalah yang telah diuraikan pada sub bab sebelumnya, maka muncul rumusan masalah sebagai berikut:

1. Bagaimana perancangan model sistem penghasil soal *vocabulary* pada *reading comprehension* TOEFL menggunakan *Learning Vector Quantization* dari artikel berita?
2. Bagaimana implementasi model sistem penghasil soal *vocabulary* pada *reading comprehension* TOEFL menggunakan *Learning Vector Quantization* dari artikel berita?
3. Bagaimana analisis terhadap hasil eksperimen dari soal yang dihasilkan oleh sistem penghasil soal *vocabulary* pada *reading comprehension* TOEFL menggunakan *Learning Vector Quantization* dari artikel berita dan teks bacaan *historical* TOEFL?

1.3 Tujuan Penelitian

Setelah diketahui rumusan masalahnya, maka tujuan dari penelitian ini adalah sebagai berikut :

1. Merancang model sistem penghasil soal *vocabulary* pada bagian *reading comprehension* TOEFL menggunakan *Learning Vector Quantization* dari artikel berita.
2. Mengimplementasikan model sistem penghasil soal *vocabulary* menggunakan *Learning Vector Quantization* dari artikel berita, menjadi API dan mengintegrasikannya dengan Aplikasi Web Penghasil Soal TOEFL yang telah dibuat pada penelitian sebelumnya.

3. Menganalisis dan validasi hasil eksperimen dari soal yang dihasilkan oleh sistem penghasil soal *vocabulary* menggunakan *Learning Vector Quantization* dari artikel berita, dengan melibatkan penilaian dari empat orang *expert*.

1.4 Manfaat Penelitian

Adapun manfaat dilakukannya penelitian ini adalah sebagai berikut:

1. Mempermudah dan mengurangi waktu yang diperlukan untuk pembuatan soal *vocabulary* TOEFL dengan memanfaatkan artikel yang aktual dari sebuah situs berita bahasa Inggris.
2. Memberikan sumber belajar bagi peserta berupa kumpulan soal yang cukup untuk mempersiapkan diri sebelum melakukan tes TOEFL.
3. Memberikan penelitian dan pengetahuan baru terhadap penggunaan *Learning Vector Quantization* serta *Natural Language Processing* dalam pembuatan soal otomatis.

1.5 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Teks narasi berita harus mempunyai *grammar* yang baik dan bebas dari teks lain selain teks berita itu sendiri.
2. Penelitian ini hanya dapat membuat jenis soal *vocabulary* TOEFL.
3. Pilihan jawaban benar yang dihasilkan belum mempertimbangkan sinonim secara semantik.

1.6 Sistematika Penulisan

Pada bagian sistematika penulisan ini akan diuraikan mengenai penjelasan tiap bab.

BAB I PENDAHULUAN

Bab ini akan menjelaskan mengenai latar belakang dari penelitian yang dilakukan, merumuskan masalah dan tujuan penelitian, juga mendefinisikan manfaat yang didapat dari penelitian ini. Dimana untuk membuat soal tes berbahasa Inggris, seseorang membutuhkan waktu, tenaga, dan biaya yang tidak sedikit. Juga tersedianya sumber belajar persiapan tes TOEFL menjadi penting,

untuk menunjang kesiapan peserta sebelum melakukan tes TOEFL. Oleh karena itu penelitian question generation ini penting untuk dilakukan.

BAB II KAJIAN PUSTAKA

Sebelum melakukan penelitian *question generation*, diperlukan pemahaman teori dan metode yang dibutuhkan untuk merancang dan membangun sistem. Adapun teori dan metode yang dibutuhkan antara lain pemahaman mengenai tes TOEFL itu sendiri, pemahaman terkait Machine Learning, Natural Language Processing, algoritma learning vector quantization, pemanfaatan lexical database wordnet, dan penelitian lain terkait Automatic Question Generation. Lebih jelasnya akan dijelaskan pada bab ini.

BAB III METODOLOGI PENELITIAN

Pada bab ini akan menjelaskan mengenai metode penelitian yang dilakukan dalam penelitian, yaitu mulai dari desain penelitian, alat dan bahan penelitian, Adapun tahapan desain penelitian terdiri dari tahapan persiapan, studi literatur, data penelitian, tahapan pembentukan model, tahapan pengembangan sistem, tahapan desain eksperimen, dan yang terakhir adalah melakukan analisis dan hasil.

BAB IV HASIL DAN PEMBAHASAN

Selanjutnya adalah melakukan langkah-langkah yang telah didesain pada bab sebelumnya. Adapun langkah-langkah yang dilakukan adalah antara lain mengumpulkan data soal yang akan menjadi data latih dan juga data artikel yang akan diekstrak untuk mendapatkan soal, yang selanjutnya akan menjadi kandidat soal jenis vocabulary pada bagian reading comprehension TOEFL. Selanjutnya melakukan proses pada data latih dan data artikel yang telah dikumpulkan, kemudian hasil dari eksperimen dilakukan analisis. Lebih jelasnya akan dijelaskan pada bab ini.

BAB V KESIMPULAN DAN SARAN

Bab terakhir yaitu bab ini akan menjelaskan kesimpulan dari hasil eksperimen yang telah dilakukan, dan menjabarkan analisis dari hasil. Juga akan

menjelaskan saran bagi peneliti selanjutnya dari penelitian yang dilakukan, agar untuk selanjutnya tidak terjadi kesalahan yang sama.

