1. Pendahuluan

Latar Belakang

Gaya berkomunikasi seseorang dipengaruhi oleh berbagai faktor, di antaranya faktor lingkungan, latar belakang pendidikan, dan kondisi emosional. Akibat dari faktor tersebut, setiap orang akan memiliki keunikan tersendiri dalam membuat sebuah kalimat. Seseorang akan memiliki perbedaan dengan orang lain dalam membuat sebuah kalimat baik itu dari susunan kata dan penggunaan kata pada kalimat yang akan dibuat. Dari kalimat yang dibuat oleh masing-masing orang, kemungkinan besar akan memiliki makna kalimat yang sama walaupun dari susunan kalimat dan penggunaan kata yang berbeda.

Hal ini juga sering terjadi pada forum *online*. Setiap pengguna memiliki cara tersendiri untuk membuat pertanyaan yang akan diajukan pada forum *online*. Akibat dari hal itu, pada forum online tersebut akan terjadi diskusi yang diajukan oleh pengguna yang memiliki makna yang sama. Akibat dari hal tersebut, sistem dari forum online tersebut mengakibatkan duplikat pertanyaan [1]. Duplikat pertanyaan akan tersimpan pada database sistem forum online tersebut secara terpisah. Dampak untuk sistem yaitu terdapat peningkatan beban pada penyimpanan dan sistem mengalami penurunan kinerja dikarenakan sistem harus mencari semua pertanyaan yang diinginkan pengguna di dalam database, termasuk pertanyaan duplikat. Hal ini dapat memperlambat kinerja sistem dan pengguna menjadi tidak puas akibat dampak tersebut. Solusi dari permasalahan tersebut adalah dengan cara mengidentifikasi kesamaan dari pertanyaan tersebut. Dengan adanya identifikasi tersebut, sistem dapat mudah mendeteksi dari pertanyaan-pertanyaan yang memiliki kesamaan makna sehingga dapat memberikan jawaban dengan cepat dan efisien. Hal ini menjadi alasan penelitian ini dilakukan yaitu melakukan analisis terhadap dua pasang pertanyaan yang memiliki makna yang sama.

Analisis tersebut dapat dilakukan dengan teknik *Natural Language Processing* (NLP). NLP adalah kumpulan teknik komputer yang termotivasi secara teoritis untuk analisis otomatis dan representasi bahasa manusia [2]. Dengan NLP, dapat dibuat metode untuk mendeteksi apakah dua pasang pertanyaan memiliki makna yang sama [3]. Salah satu metode yang dapat digunakan adalah *Convolutional Neural Network* (CNN) dengan cara menghitung nilai kesamaan antara vektor representasi [4].

Metode CNN sudah banyak digunakan untuk penelitian kesamaan teks dan mendapatkan hasil akurasi yang cukup besar [4]. Penerapan metode CNN dalam kasus *question similarity* ini sudah digunakan pada penelitian yang dilakukan oleh Dasha et al. dengan judul *Detecting Semantically Equivalent Questions in Online User Forums* [4]. Pada penelitian tersebut, para peneliti membuat sistem mendeteksi kesamaan semantik menggunakan CNN dan menggunakan data dari website *Ask Ubuntu Community Questions and Answers* dan *Meta Stack Exchange* pada tahun 2015. Data tersebut merupakan pertanyaan dari komunitas pengguna dan pengembang ubuntu. Para peneliti mengimplementasikan CNN dengan cara mengubah kata menjadi *word embedding* menggunakan kumpulan data tanpa label dan menerapkan CNN untuk membuat representasi vektor terdistribusi untuk pasangan pertanyaan. Sebagai pembanding dari metode CNN, mereka memilih metode lain yaitu *Support Vector Machine* (SVM) sebagai metode pembanding. Hasil penelitian tersebut menunjukkan bahwa CNN menghasilkan *valid accuracy* 93,4% dan *test accuracy* 92,9% [4] dimana hasil tersebut lebih besar dibandingkan menggunakan metode SVM yang menghasilkan *valid accuracy* 85,5% dan *test accuracy* 82,4%.

Rumusan masalah pada penelitian ini adalah bagaimana sistem identifikasi kesamaan dari pertanyaan-pertanyaan dibuat dengan menerapkan metode CNN, dan bagaimana hasil evaluasi sistem identifikasi kesamaan dari pertanyaan-pertanyaan menggunakan metode CNN.

Topik dan Batasannya

Topik pada penelitian Tugas Akhir (TA) ini adalah membuat sebuah model untuk mengidentifikasi kesamaan dari sepasang pertanyaan pada forum komunitas online menggunakan metode CNN. Metode CNN digunakan pada penelitian karena beberapa penelitian sebelumnya dengan topik yang sama menggunakan metode CNN mendapatkan hasil lebih baik dibandingkan dengan metode lainnya [4]. CNN dapat menangkap fitur lokal pada data, yaitu frase dan kata penting yang ada pada kalimat.

Batasan masalah dalam penelitian TA ini adalah jumlah data yang akan digunakan untuk penelitian ini hanya memiliki 4.752 data dimana data tersebut relatif lebih sedikit dibandingkan dengan jumlah data yang digunakan untuk topik sejenis. Proses pelabelan data dilakukan secara manual.

Tujuan

Tujuan yang ingin didapatkan dari penelitian ini adalah membuat sebuah sistem yang dapat digunakan untuk mengidentifikasi pertanyaan-pertanyaan apakah memiliki kesamaan dari pertanyaan satu sama lain menggunakan metode CNN. Untuk mengetahui efektivitas penggunaan metode CNN pada sistem identifikasi kesamaan pertanyaan, dilakukan evaluasi untuk mendapatkan nilai *accuracy*, *precision*, *recall*, dan *F1-Score*.