METODE "BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS" (BERT) DALAM PERHITUNGAN *SEMANTIC SIMILARITY*

SKRIPSI

Untuk memenuhi sebagian persyaratan
Mencapai gelar Sarjana
pada Program Studi Informatika

disusun oleh

Dhimas Yoga Pratama

17.11.0972

PROGRAM SARJANA
PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2020

METODE "BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS" (BERT) DALAM PERHITUNGAN *SEMANTIC SIMILARITY*

SKRIPSI

Untuk memenuhi sebagian persyaratan
Mencapai gelar Sarjana
pada Program Studi Informatika

disusun oleh

Dhimas Yoga Pratama

17.11.0972

PROGRAM SARJANA
PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2020

PERSETUJUAN

SKRIPSI

METODE "BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS" (BERT) DALAM PERHITUNGAN *SEMANTIC SIMILARITY*

yang dipersiapkan dan disusun oleh

Dhimas Yoga Pratama

17.11.0972

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 25 November 2020

Dosen Pembimbing,

Mardhiya Hayaty, S.T., M.Kom.
NIK. 190302108

PENGESAHAN

SKRIPSI

METODE "BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS" (BERT) DALAM PERHITUNGAN *SEMANTIC SIMILARITY*

yang dipersiapkan dan disusun oleh

Dhimas Yoga Pratama

17.11.0972

telah dipertahankan di depan Dewan Penguji
pada tanggal 18 November 2020

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Ainul Yaqin, M.Kom **NIK. 190302255**	_____
Bayu Setiaji, M.Kom **NIK. 190302216**	_____
Mardhiya Hayaty, S.T., M.Kom. **NIK. 190302108**	_____

Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 18 November 2020

DEKAN FAKULTAS ILMU KOMPUTER

Krisnawati, S.Si, M.T.
NIK. 190302038

PERNYATAAN

Saya yang bertandatangan di bawah ini menyatakan bahwa skripsi ini merupakan karya saya sendiri (ASLI) dan isi pada skripsi ini tidak terdapat karya yang pernah diajukan oleh orang lain untuk memperoleh gelar akademis di suatu institusi pendidikan tinggi manapun, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis dan/atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Segala sesuatu yang terkait dengan masalah dan karya yang telah dibuat adalah menjadi tanggungjawab saya pribadi.

Yogyakarta, 25 November 2020

Dhimas Yoga Pratama

NIM 17.11.0972

MOTTO

"*Sesungguhnya bersama kesulitan pasti ada kemudahan. Maka apabila engkau telah selesai (dari suatu urusan), tetaplah bekerja keras (untuk urusan yang lain)*"

(QS 94:6-7)

"*The best thing you can ever do is follow your dreams*"

(James Owen Sullivan)

PERSEMBAHAN

Alhamdulillahirabbil‘alamin, segala puji bagi Allah SWT atas izin dan rahmat-Nya serta doa, usaha, juga kerja keras skripsi ini dapat diselesaikan dengan baik. Pada penyusunan skripsi ini penulis tentu dibantu oleh berbagai pihak, untuk itu dengan ini penulis mempersembahkan ucapan terimakasih sedalam-dalamnya kepada :

1. Ibu saya Dini Sintya Anggraeni dan bapak saya Widayanto yang senantiasa memberikan doa serta dukungan.
2. Segenap keluarga yang telah hadir sebagai lingkungan dimana penulis dididik, tumbuh dan berkembang hingga dewasa.
3. Dosen pembimbing saya Ibu Mardhiya Hayaty, S.T., M.Kom. yang telah membimbing saya dengan sabar dan baik dari awal hingga skripsi ini selesai.
4. Dosen Universitas AMIKOM Yogyakarta yang telah memberikan banyak ilmu yang bermanfaat selama perkuliahan.
5. Alde Satriayu Putri Nadeak yang selalu mendampingi, memberikan dukungan, motivasi serta doa.
6. Rekan sepenelitian saya Anjas, Andri, Lutfi, Dimas Midyan yang selalu membantu, menemani, dan berbagi ilmu selama pengerjaan skripsi ini.
7. Segenap teman-teman IF-01 khususnya Aldo, Parmin, Zul, Roy, Irfan, Dedi, Dendy, Dandy, Jeki, Fariz, dan Yoga yang telah menemani, berbagi canda tawa, duka, dan juga sejumlah uang selama 3 tahun masa perkuliahan.
8. Pihak lain yang tidak dapat saya sebutkan satu persatu.

KATA PENGANTAR

Penulis panjatkan puji dan syukur kehadirat Allah SWT atas izin, berkah, dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik yang berjudul "**METODE *BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMERS* (BERT) DALAM PERHITUNGAN *SEMANTIC SIMILARITY***" disusun sebagai salah satu syarat utama untuk menyelesaikan program sarjana di Universitas AMIKOM Yogyakarta. Penyelesaian skripsi ini juga tidak lepas dari bantuan berbagai pihak, karena itu pada kesempatan ini penulis ingin menyampaikan rasa hormat dan terimakasih kepada :

1. Prof. Dr. M. Suyanto, MM. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Ibu Krisnawati, S.Si., M.T. selaku Dekan Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta.
3. Bapak Sudarmawan, M.T. selaku Ketua Program Studi Informatika Universitas AMIKOM Yogyakarta.
4. Ibu Mardhiya Hayaty, S.T., M.Kom. selaku dosen pembimbing yang telah memberikan bimbingan dengan baik selama penyusunan skripsi ini.
5. Bapak Ainul Yaqin, M.Kom. dan Bapak Bayu Setiaji, M.Kom. selaku dosen penguji.

Penulis menyadari bahwa skripsi ini masih terdapat kekurangan. Maka penulis menerima segala kritik dan saran yang membangun serta teguran dari semua pihak. Semoga skripsi ini dapat bermanfaat bagi penulis, pembaca, dan perkembangan *Deep Learning* di Indonesia. Atas saran dan kritik penulis ucapkan terimakasih.

Yogyakarta, 25 November 2020

Dhimas Yoga Pratama

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Quora adalah situs forum *online* yang digunakan oleh banyak pengguna dari seluruh dunia untuk melakukan tanya jawab mengenai topik tertentu. Mengingat banyaknya pengguna yang menggunakan layanan ini dan aktif dalam melakukan tanya jawab menyebabkan setiap pengguna menanyakan pertanyaan dengan makna yang sama namun dengan kata yang berbeda, inilah yang menyebabkan adanya masalah duplikasi pertanyaan. Oleh sebab itu diperlukan metode yang dapat mengidentifikasi adanya duplikasi dari pertanyaan yang di-*input*kan pengguna.

Penelitian ini menggunakan sebuah metode dalam *deep learning* yaitu *Bidirectional Encoder Representation from Transformer* atau BERT sebagai metode untuk membuat *word embedding*. Untuk memprediksi probabilitas kemiripan pertanyaan digunakan metode *logistic regression*. Pertanyaan yang diidentifikasi akan memiliki 2 label yaitu *duplicate* dan *not_duplicate*.

Proses permodelan dilakukan sebanyak 8 *epochs*. Data yang digunakan merupakan judul pertanyaan yang diambil dari situs Quora dengan teknik *scraping*. Total pasangan judul pertanyaan adalah sebanyak 10.000 pasang dengan pembagian data latih sebanyak 70%, data evaluasi sebanyak 10%, dan data uji sebanyak 20%. Dari hasil pengujian prediksi probabilitas kemiripan didapatkan nilai akurasi sebesar 99% .

Kata Kunci : Quora, *Deep Learning*, BERT, *Word Embedding*, *Logistic Regression.*

ABSTRACT

Quora is an online forum used by many users from all over the world to ask and answer questions about certain topics. Given the large number who use this service and are active in questioning and answering, it causes each user to ask the same question(s) but in different way, this then causes the duplication of questions problem. Therefore, we need a method that can identify the duplication of user-inputted questions.

This research use a method in Deep Learning namely Bidirectional Encoder Representation from Transformers or BERT as a method to create word embedding. Logistic regression is then used to predict the probability of the similarity of the questions. The identified questions will have 2 labels namely duplicate and not_duplicate.

The modelling process is carried out in 8 epochs. The data used is the title of the questions taken from Quora site using scraping technique. The total number of pairs of questions are 10.000 pairs, which then splitted into training data as much as 70%, evaluation data as much as 10%, and training data as much as 20%. From the testing of probability of the similarity prediction obtained accuracy in the amount of 99%.

Keywords: *Quora, Deep Learning, BERT, Word Embedding, Logistic Regression.*