IDENTIFIKASI DUPLIKAT PERTANYAAN MENGGUNAKAN METODE *GLOBAL VECTOR* (GLOVE)

SKRIPSI

disusun oleh

Yohanes Eudes Anjas Susetya

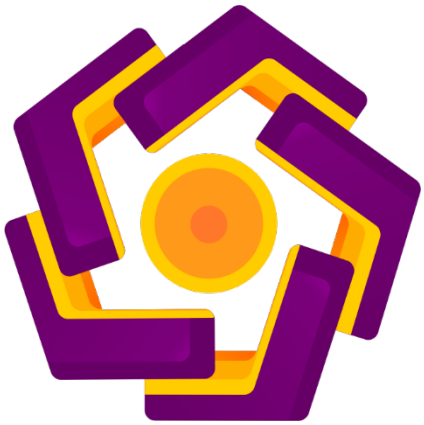
17.11.0922

PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2020

IDENTIFIKASI DUPLIKAT PERTANYAAN MENGGUNAKAN METODE *GLOBAL VECTOR* (GLOVE)

SKRIPSI

untuk memenuhi sebagian persyaratan
mencapai gelar Sarjana
pada Program Studi Informatika



disusun oleh

Yohanes Eudes Anjas Susetya

17.11.0922

PROGRAM SARJANA
PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2020

PERSETUJUAN

SKRIPSI

IDENTIFIKASI DUPLIKAT PERTANYAAN MENGGUNAKAN METODE *GLOBAL VECTOR* (GLOVE)

yang dipersiapkan dan disusun oleh

Yohanes Eudes Anjas Susetya

17.11.0922

telah disetujui oleh Dosen Pembimbing Skripsi

pada tanggal 23 November 2020

Dosen Pembimbing,

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

NIK. 190302108

PENGESAHAN

SKRIPSI

IDENTIFIKASI DUPLIKAT PERTANYAAN MENGGUNAKAN METODE *GLOBAL VECTOR* (GLOVE)

yang dipersiapkan dan disusun oleh

Yohanes Eudes Anjas Susetya

17.11.0922

telah dipertahankan di depan Dewan Penguji
pada tanggal 18 November 2020

Susunan Dewan Penguji

Nama Penguji	Tanda Tangan
Barka Satya, M.Kom. **NIK. 190302126**	_____
Nuraini, M.Kom. **NIK. 190302066**	_____
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, 21 November 2020

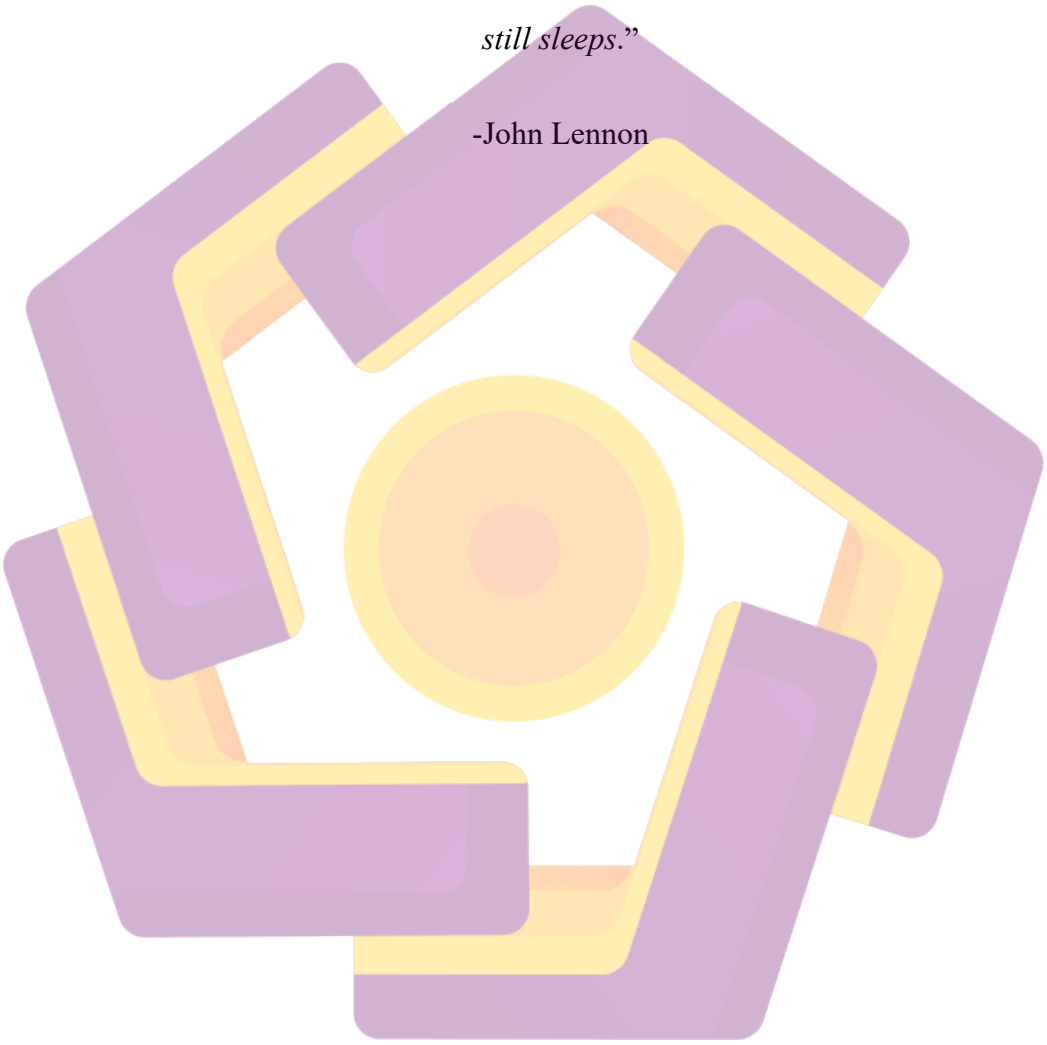
Yohanes Eudes Anjas Susetya

NIM 17.11.0922

MOTTO

"*When you do something noble and beautiful and nobody noticed, do not be sad. For the sun every morning is a beautiful spectacle and yet most of the audience still sleeps.*"

-John Lennon



PERSEMBAHAN

Puji Tuhan berkat kerja keras serta doa, skripsi ini dapat diselesaikan dengan baik. Segala puji dan syukur kepada Tuhan Yang Maha Esa yang tiada henti memberikan berkat serta kekuatan untuk dapat menyelesaikan skripsi. Dengan ini saya mempersembahkan skripsi ini kepada semua pihak yang terlibat secara langsung atau tidak langsung, yaitu untuk :

1. Kedua orang tua saya (Theopilus Suryantoro Budi Susilo dan Theresia Wagiyem).
2. Adik saya (Wilfridus Bimo Ardiyantoro) dan bude (Yustina Widaryanti).
3. Dosen pembimbing saya Ibu Mardhiya Hayaty, S.T., M.Kom., yang telah sabar dalam membimbing saya dari awal sampai skripsi ini selesai.
4. Dosen Universitas AMIKOM Yogyakarta yang telah memberikan banyak ilmu selama perkuliahan berlangsung.
5. Teman sepenelitian (Dhimas, Lutfi, Andri, Bangdim) yang selalu membantu, menemani dan memberikan dukungan kepada saya untuk dapat menyelesaikan skripsi ini.
6. Teman sekelas 17-IF-01 yang telah memberikan dukungan selama perkuliahan.
7. Teman sorak-sorai yang senantiasa memberikan semangat (Devita, Gilang, Dendy, Bagas, Lordjek, Pandu, Aldo, Bapok, Roy, Arfian dan Dito).
8. Tim valorant penghuni *iron* yang senantiasa memberikan semangat dan petunjuk dalam setiap *round game* (Bram, Emo, Parmin, Fariz).

KATA PENGANTAR

Penulis panjatkan puji dan syukur kehadirat Tuhan Yang Maha Esa atas berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi dengan baik yang berjudul "**IDENTIFIKASI DUPLIKAT PERTANYAAN MENGGUNAKAN METODE *GLOBAL VECTOR* (GLOVE)**" disusun sebagai salah satu syarat utama untuk menyelesaikan program sarjana pada Universitas AMIKOM Yogyakarta. Penyelesaian skripsi ini juga tidak lepas dari bantuan berbagai pihak, karena itu pada kesempatan ini penulis ingin menyampaikan rasa hormat dan terima kasih 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 selalu bijaksana dalam memberikan bimbingan serta nasehat dan waktunya selama pengerjaan skripsi ini.
5. Bapak Barka Satya, M.Kom. dan Ibu Nuraini, M.Kom. selaku dosen penguji. Terimakasih atas segala saran yang diberikan selama pengujian untuk memperbaiki penelitian menjadi lebih baik lagi.

Penulis menyadari bahwa skripsi ini masih ada kekurangan. Maka, penulis menerima segala kritik dan saran yang membangun serta teguran dari semua pihak. Semoga skripsi ini bisa bermanfaat baik bagi penulis serta pembaca. Atas saran dan kritik penulis ucapkan terima kasih.

Yogyakarta, 23 November 2020

Yohanes Eudes Anjas Susetya

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Forum diskusi *online* merupakan tempat untuk berkumpul pengguna internet yang di dalamnya kita bisa berbagi informasi dan berdiskusi antar pengguna forum mengenai suatu topik tertentu. Mengingat semakin bertambahnya pengguna internet saat ini juga membuat forum diskusi semakin ramai dan aktif dalam melakukan tanya jawab. Dengan banyaknya pengguna menyebabkan setiap pengguna di dalam forum menanyakan pertanyaan dengan maksud yang sama namun dengan kata yang berbeda. Oleh sebab itu penelitian ini menggunakan data pertanyaan forum tanya jawab untuk melakukan identifikasi duplikat pertanyaan yang diperoleh dari situs web Quora Indonesia.

Pada penelitian ini data diklasifikasikan menggunakan *Bidirectional Long Short Term Memory* atau Bi-LSTM. Pertanyaan yang diidentifikasi akan memiliki 2 label yaitu *duplicate* dan *not_diplicate*. Untuk merepresentasikan kata ke dalam vektor, penelitian ini menggunakan *word embedding*. Penelitian ini bertujuan untuk mengetahui berapa hasil akurasi dari *word embedding* yang digunakan dalam identifikasi duplikat pasangan pertanyaan menggunakan Bahasa Indonesia.

Metode *word embedding* yang digunakan adalah *Global Vector* (GloVe). Dari percobaan *word embedding* metode *Global Vector* (GloVe) yang dilakukan dengan pembagian data *training* 80% dan data *test* 20% dari total jumlah dataset 10000 mendapatkan hasil akurasi 95% dengan presisi 96%, *recall* 93% dan *f1-score* 94%.

Kata-kunci: Quora, Identifikasi Pertanyaan, *Word Embedding, GloVe, Bi-LSTM*

ABSTRACT

An online discussion forum is a place for internet users to gather in which we can share information and discuss among forum users on a particular topic. Considering the increasing number of internet users nowadays it also makes discussion forums more crowded and active in conducting questions and answers. With so many users it causes every user in the forum to ask questions with the same intent but with different words. Therefore this study uses question and answer forum question data to identify duplicate questions obtained from the Quora Indonesia website.

In this study, the data were classified using Bidirectional Long Short Term Memory or Bi-LSTM. Identified questions will have 2 labels namely duplicate and not_diplicate. To represent words into vectors, this study uses word embedding. This study aims to determine the accuracy results of the word embedding used in the identification of duplicate question pairs using Indonesian.

The word embedding method used is Global Vector (GloVe). From the word embedding experiment with the Global Vector (GloVe) method, which was carried out by sharing 80% training data and 20% test data of the total number of 10000 datasets, the results obtained 95% accuracy with 96% precision, 93% recall and 94% f1-score.

Keywords: *Quora, Identification questions pairs, Word Embedding, GloVe, Bi-LSTM*