BAB I
PENDAHULUAN

1.1 Latar Belakang

Teknologi informasi adalah salah satu contoh produk teknologi yang pada saat ini terus berkembang dari hari ke hari baik itu informasi dari media cetak maupun elektronik yang menyajikan informasi bentuk tulisan, suara maupun gambar. *Website* adalah salah satu teknologi informasi yang mampu memberikan informasi menjadi lebih mudah. *Question Answering* berkembang cepat dalam dunia internet *(online)* khususnya *website*. *Website* forum diskusi *online* adalah *platform* untuk berkumpul, berbagi informasi dan berdiskusi antar pengguna mengenai suatu topik tertentu. Pengguna pada forum diskusi *online* dapat bertanya mengenai suatu topik, kemudian pengguna lain yang ahli mengenai topik yang ditanyakan menjawab pertanyaan tersebut [1].

Makna disetiap kata *question answer* dalam forum diskusi *online* masih belum sesuai. Terkadang pertanyaan yang diajukan mengandung kata (data teks) yang bias dan sama, juga makna kata yang masih belum tepat. Berawal dari data teks atau kata yang tidak terstruktur, para peneliti menggunakan metode *Natural Language Processing* untuk dapat memecahkan masalah. *Natural Language Processing* adalah sebuah metode pembentukan model komputasi dengan perantara bahasa alami yang berupaya memahami bahasa alami manusia dengan segala aturan gramatika dan semantiknya, serta mengubah bahasa tersebut menjadi representasi formal yang dapat diproses oleh komputer [2].

Permasalahan kata (data teks) dalam forum diskusi yang umumnya muncul adalah ketika salah satu atau beberapa pengguna lainnya mengajukan pertanyaan dengan maksud yang sama. Hal ini akan menjadi sebuah permasalahan dalam efisiensi sebuah *website* tanya jawab atau forum diskusi serta penggunaan ruang penyimpanan yang besar untuk informasi yang sama. Pertanyaan-pertanyaan dengan maksud yang sama akan tersebar di berbagai forum diskusi, padahal jika digabungkan akan lebih baik dan memaksimalkan jawaban oleh pengguna lain [3].

Penulis mengajukan penelitian identifikasi dengan *word embedding (Global Vector)* sehingga nantinya kata (data teks) yang tersedia dalam forum diskusi *online* dapat diolah, dianalisa dan dicari informasi penting di dalamnya untuk kemudian dapat dilakukan proses vektor sehingga dapat mengetahui informasi pada duplikasi pertanyaan berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan dengan penjabaran latar belakang di atas, diperoleh rumusan masalah yaitu berapakah hasil akurasi penggunaan *word embedding* dengan metode *Global Vector* dalam identifikasi duplikat pertanyaan dengan topik seputar "Teknologi Informasi dan Komunikasi" dari situs Quora Indonesia.

1.3 Batasan Masalah

Agar Penelitian lebih terfokus dan tidak menyimpang, maka penulis menetapkan batasan masalah yakni data yang diambil dari situs Quora Indonesia dan hanya terbatas pada topik seputar "Teknologi Informasi dan Komunikasi".

Bahasa pertanyaan yang digunakan adalah Bahasa Indonesia baku dan tidak terdapat singkatan.

1.4 Maksud dan Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maksud dan tujuan dilakukannya penelitian ini yaitu mengimplementasikan *word embedding* pada forum *question answer* untuk mengetahui nilai akurasi Global Vector (GloVe) dalam identifikasi duplikat pertanyaan dalam Bahasa Indonesia.

1.5 Manfaat Penelitian

Adapun manfaat penelitian adalah mengetahui tingkat akurasi untuk identifikasi duplikat pertanyaan dengan penerapan *Global Vector* (GloVe).

1.6 Metode Penelitian

Berikut metode penelitian yang digunakan dalam proses pengerjaan penelitian ini dijelaskan sebagai berikut :

1.6.1 Pengumpulan Data

Dalam penelitian ini pengumpulan data yang digunakan dengan mengambil data atau *scraping* dari situs Quora Indonesia sebanyak 10000 pasang pertanyaan yang tertuju pada topik pembahasan seputar "Teknologi Informasi dan Komunikasi". Semua data yang diambil berupa teks. Teks terdiri dari bentuk yang tidak terstruktur sehingga harus diubah menjadi data yang diolah yang nantinya digunakan dalam proses pengujian penelitian.

1.6.2 *Annotation*

Setelah data yang di*scraping* dari situs Quora Indonesia selesai, kemudian dianalisis dengan memberikan label pada setiap pertanyaan. Label dalam penelitian ini diberinama "*is_duplicate*" yang digunakan untuk menentukan pertanyaan tersebut duplikat atau tidak duplikat. Label penelitian ini menggunakan label 1 untuk pertanyaan yang duplikat sedangkan 0 untuk pertanyaan yang tidak duplikat. Tahapan *Text annotation* membantu untuk memvisualisasikan teks atau kata penting melalui *computer vision* saat pelabelan teks, teks disorot dan metadata juga ditambahkan di setiap kata yang penting untuk diintegrasikan ke dalam pemrosesan Bahasa kemudian memberikan pembelajaran pada model yang akan dibentuk di tahap pelatihan data.

1.6.3 *Preprocessing* Data

Preprocessing data pada penelitian ini melalukan perubahan data teks yang mentah agar siap diolah dalam pengujian. *Preprocessing* meliputi data *cleaning, case folding, filtering* (*stopword removal*) dan *tokenization*.

1.6.4 Ekstraksi Fitur

Ekstraksi Fitur atau pembangunan vektor pada tahapan penelitian ini dilakukan untuk mengubah kata dari kata-kata menjadi numerik karena komputer hanya bisa membaca sebuah angka. Ekstraksi fitur atau pembangunan vector memakai ekstraksi fitur *word embedding* bagian dari ekstraksi fitur, pada pembangunan vector penelitian ini menggunakan metode *Global Vector* (GloVe).

1.6.5 Implementasi Algoritma Klasifikasi

Pada tahapan ini dilakukan proses klasifikasi data yang telah melalui tahap ekstraksi fitur. Implementasi algoritma klasifikasi yang akan digunakan dalam penelitian ini adalah *Bidirectional - Long Short Term Memory bidirectional*.

1.6.6 Evaluasi

Hasil dari implementasi selanjutnya dilakukan proses pengujian. Akan dilihat keberhasilan algoritma untuk mengklasifikasikan pertanyaan dengan data *test* yang dilakukan. Dari hal tersebut dapat dilihat keakurasian atau performansi dari algoritma yang digunakan terhadap data yang ada. Untuk mengetahui akurasi dari perhitungan algoritma digunakan tabel perhitungan *confusion matrix* antara label sebenarnya dengan label hasil pengujian data pertanyaan setelah dilakukan pengujian. Label yang sama dapat dikatakan benar dan label yang tidak sama dapat dikatakan salah. Maka dari bagian akurasi dapat diketahui melalui kebenaran labelnya.

1.7 Sistematika Penulisan

Sistematika penulisan dibuat untuk mempermudah penulis dalam peyusunan skripsi. Adapun sistematika penulisan ini dikelompokan kedalam beberapa bab. Setiap bab diuraikan sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini menjelaskan tentang dasar penelitian, yang berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metode penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Pada bab ini berisi tinjauan pustaka yang mirip dengan penelitian ini. Pada bab ini juga berisi tentang landasan teori yang mendukung dalam penelitian ini.

BAB III METODE PENELITIAN

Pada bab ini berisi tentang alur dari penelitian yang berupa perancangan perangkat dan bahan yang akan digunakan.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang hasil dari tahapan penelitian yang dilakukan secara menyeluruh termasuk hasil dari pengujian.

BAB V PENUTUP

Bagian terakhir dari penelitian yang berisi tentang kesimpulan dan saran untuk memperbaiki kekurangan yang ada pada penelitian ini.