BAB I

PENDAHULUAN

1.1 Latar Belakang

Natural Language Processing (NLP) merupakan suatu bidang pada *Artificial Intelligence* (AI) yang berfokus agar suatu komputer dapat memproses konteks pada suatu kalimat seperti halnya pada manusia. Salah satu kegunaan NLP adalah untuk memprediksi kata selanjutnya yang diketik, atau yang lebih dikenal sebagai *next-word prediction.* Agar sistem *next-word prediction* dapat dibuat, sistem harus mengingat kata-kata sebelumnya yang diketik oleh pengguna agar prediksi yang diberikan dapat sesuai dengan konteks kalimat tersebut. Hal ini dapat dilakukan dengan menggunakan arsitektur *Recurrent Neural Network* (RNN).

Dalam pembuatan sistem *next-word prediction,* LSTM kerap lebih digunakan daripada RNN dikarenakan LSTM dapat menyimpan semua kata-kata yang dianggap bermanfaat untuk memprediksikan kata selanjutnya pada sebuah teks dan dapat melupakan kata-kata yang dianggap tidak terlalu bermanfaat. Hal ini terlihat dari perbandingan performa RNN dan LSTM pada tugas pemodelan bahasa, dimana akurasi meningkat kurang lebih sebesar 8% [1]. Penulis memutuskan untuk menggunakan LSTM dibandingkan arsitektur model berbasis RNN lainnya seperti performa LSTM melewati GRU sewaktu digunakan untuk membuat model menggunakan dataset yang kompleks.[2]

Salah satu aspek yang penting dalam pembuatan *next-word prediction* adalah personalisasi. Personalisasi diperlukan sehingga prediksi yang diberikan sesuai dengan pemilihan kata pengguna. Akan tetapi, melakukan personalisasi pada umumnya berarti pengguna harus mengorbankan privasi. Agar personalisasi dapat dilakukan, beberapa data personal harus dikirim ke server sehingga model yang sesuai dengan pengguna tersebut dapat dilatih. Data personal tersebut dapat berisi hal-hal yang sensitif seperti nomor kartu kredit, alamat rumah, dan identitas keluarga. Mengingat bahwa terdapatnya

ketidakpercayaan yang luas pada masyarakat mengenai cara data dikoleksi oleh perusahaan [1], sistem *next-word prediction* yang dibuat harus dapat menjamin masyarakat bahwa sistem tersebut dapat menjaga privasi.

Terdapat beberapa cara yang biasa digunakan untuk menjaga privasi pengguna dalam sistem *next-word prediction.* Salah satunya adalah *on-device offline learning,* dimana model disimpan pada perangkat sehingga pelatihan dapat dilakukan secara lokal saja. Dengan membuat sistem *next-word prediction* dengan metode tersebut, privasi data dapat dijaga dikarenakan tidak terdapat data apa pun yang dikirimkan kembali ke server.

Metode lain yang dapat digunakan untuk menjaga privasi pengguna adalah *Federated Learning* (FL) [4], dimana model dilatih pada perangkat telepon genggam sehingga hanya parameter model yang dikirim ke server. Walaupun hal ini lebih aman daripada mengirimkan data personal ke server secara langsung, sebuah studi yang menganalisis aspek privasi pada FL menunjukkan bahwa desain sistem FL tidak sepenuhnya kebal dari masalah privasi [5]. Model parameter yang dikirim dari lokal ke server berisi informasi sensitif dapat dieksploitasi untuk mendapatkan informasi personal pengguna. Sistem FL juga merupakan target dari serangan seperti *model poisoning* & *inference attack* yang dapat mengakibatkan kebocoran privasi.

Penelitian ini akan membuat sistem *next-word prediction* secara *on-device* dan dilatih dengan menggunakan *offline learning* dengan menggunakan algoritma LSTM. Penulis juga akan melakukan riset tersebut dengan menggunakan Bahasa Indonesia dikarenakan masih sedikitnya jumlah riset yang dilakukan dalam Bahasa Indonesia pada bidang NLP. Walaupun sudah terdapat beberapa riset yang mengimplementasikan sistem *offline learning*, sejauh pengetahuan penulis, penulis menemukan bahwa belum ada implementasi sistem tersebut secara *On-Device* dalam Bahasa Indonesia. Penulis juga mengenalkan metode cache yang dapat digunakan untuk mengatasi kata-kata yang belum terdapat dalam vocabulary tokenizer.

Dalam penelitian ini, penulis akan menggunakan data yang diambil dari Twitter. Penulis memutuskan untuk menggunakan data-data dari Twitter untuk

pembuatan model karena cara penulisan yang digunakan pada Twitter kerap bersifat lebih informal dan lebih kerap digunakan dalam percakapan sehari-hari dibandingkan dengan data yang didapatkan dari Wikipedia sehingga lebih cocok digunakan untuk sistem yang akan memprediksikan kata selanjutnya sewaktu menggunakan bahasa yang lebih informal. Penulis juga akan melakukan optimasi terhadap model dengan tujuan untuk membuat model lebih efektif dalam melakukan prediksi dan pelatihan ulang. Beberapa optimasi yang dilakukan berupa membersihkan data-data Twitter, melakukan model kompresi dengan teknik pruning, dan mengubah learning rate model yang telah dilatih sehingga dapat lebih efektif digunakan untuk memberikan prediksi yang sesuai. Untuk menentukan apakah model sudah efektif atau belum, penulis akan menggunakan metrik top-3 eff yang akan memberikan hasil pengukuran sesuai dengan berapa banyak huruf yang harus diketik oleh pengguna sebelum pengguna mendapatkan prediksi yang diinginkan.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang sudah dipaparkan di atas, rumusan masalah pada penelitian ini terdiri dari beberapa poin yaitu

1.2.1. Bagaimana data Twitter dapat dibersihkan sehingga data dapat digunakan untuk melakukan pelatihan model?

1.2.2. Apa saja dampak yang dialami model sewaktu dilakukan pruning dan diubah menjadi bentuk tflite?

1.2.3. Apa learning rate yang paling cocok digunakan untuk melakukan personalisasi model secara *on-device?*

1.2.4. Apakah model yang telah dilakukan personalisasi dengan menggunakan dataset dari logchat penulis mempengaruhi nilai top-3 eff sewaktu dilakukan pengetesan menggunakan dataset baru?

1.2.5. Bagaimana model dapat mengatasi kata-kata yang tidak terdapat dalam vocabulary tokenizer?

1.2.6. Berapa banyak waktu dan sumber daya yang termakan untuk melakukan inferensi dan personalisasi model secara *on-device* dengan

menggunakan model yang telah di-pruning dan dengan model yang belum di-pruning?

1.3 Batasan Penelitian

Batasan masalah dari penelitian ini adalah sebagai berikut:

1.3.1. Sistem yang dirancang hanya dites dengan menggunakan beberapa perangkat telepon genggam berbasis Android, sehingga performa model bisa berbeda pada model perangkat telepon genggam lainnya

1.3.2. Sistem yang dirancang hanya menggunakan dataset yang tersedia secara publik, seperti dari Twitter, untuk mengukur performa sehingga performa dapat berubah dengan dataset lainnya

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengimplementasi sistem *next-word Prediction* menggunakan LSTM pada *on-device* yang akan dilatih secara *offline training* dan mengukur performa sistem pada model sebelum dan sesudah personalisasi

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1.5.1. Memajukan riset dalam bidang NLP pada Bahasa Indonesia

1.5.2. Meningkatkan kualitas *next-word prediction* pada Bahasa Indonesia