

SISTEM PERINGKAS OTOMATIS ABSTRAKTIF DENGAN MENGGUNAKAN RECURRENT NEURAL NETWORK

Kuncoro Yoko, Viny Christanti M, Janson Hendryli

Teknik Informatika Universitas Tarumanagara

Jl. S. Parman No.1, Blok R Lantai XI, Jakarta 11440 Indonesia

Email: kuncoroyoko@gmail.com, viny@untar.ac.id, jansonh@fti.untar.ac.id

Abstrak

Peringkasan teks otomatis berusaha untuk membuat teks lebih pendek sambil mempertahankan maknanya. Paper ini menggunakan Recurrent Neural Network (RNN) untuk membuat ringkasan dari teks Bahasa Indonesia. Korpus yang digunakan didapatkan dari situs www.detik.com dan www.kompas.com. Word2vec digunakan untuk membuat word embedding dari korpus kemudian data set dilatih dengan RNN untuk membuat sebuah model. Model ini digunakan untuk membuat berita. Model terbaik dicari dengan mengganti ukuran hidden state word2vec dan RNN. Pengujian sistem dan Q&A Evaluation digunakan untuk mengevaluasi model. Pengujian sistem menunjukkan model dengan 6457 data set, word2vec berukuran 200, dan hidden state RNN berukuran 256 memberikan akurasi terbaik yaitu 99.8810%. Model ini kemudian dievaluasi dengan Q&A Evaluation. Q&A Evaluation menunjukkan model memberikan akurasi sebesar 46.65%.

Kata kunci— *Information Retrieval, Peringkasan Teks Otomatis, Peringkasan Teks Abstraktif, Recurrent Neural Network, RNN Encoder-Decoder*

Abstract

Abstractive Text Summarization try to creates a shorter version of a text while preserve its meaning. We try to use Recurrent Neural Network (RNN) to create summaries of Bahasa Indonesia text. We get corpus from Detik dan Kompas site news. We used word2vec to create word embedding from our corpus then train our data set with RNN to create a model. This model used to generate news. We search the best model by changing word2vec size and RNN hidden states. We use system evaluation and Q&A Evaluation to evaluate our model. System evaluation showed that model with 6457 data set, 200 word2vec size, and 256 RNN hidden states gives best accuracy for 99.8810%. This model evaluated by Q&A Evaluation. Q&A Evaluation showed that the model gives 46.65% accuracy.

Key words—*Information Retrieval, Automatic Text Summarization, Abstractive Text Summarization, Recurrent Neural Network, RNN Encoder-Decoder*

1. PENDAHULUAN

Kegiatan membaca adalah salah satu hal yang sering dilakukan oleh manusia. Perkembangan teknologi menyebabkan melimpahnya berita, artikel, dan bacaan lainnya yang dapat dicari. Hal ini dapat menghabiskan banyak waktu untuk membaca bacaan-bacaan tersebut. Ringkasan adalah suatu cara yang efektif untuk menyajikan suatu karangan yang panjang dalam waktu yang singkat. Oleh sebab itu, ringkasan dapat mempermudah untuk mengambil inti dari suatu bacaan [1].

Peringkasan terdiri dari dua tipe yaitu ekstraktif dan abstraktif. Ekstraktif meringkas suatu dokumen dengan memilih sebagian dari kalimat yang ada dalam dokumen asli. Abstraktif melakukan peringkasan dengan cara menginterpretasi teks asal melalui proses transformasi suatu kalimat asli [2].

Peringkasan otomatis sudah banyak dibuat untuk Bahasa Indonesia. Contoh dari peringkasan otomatis yang telah dibuat adalah peringkasan dokumen Bahasa Indonesia yang dibuat oleh Achmad Ridok bersifat ekstraktif. Peringkasan ini dibuat dengan metode *Non-Negative Matrix Factorization* [2]. Peringkasan otomatis berbahasa Indonesia banyak dibuat dengan metode ekstraktif. Oleh karena itu, penelitian ini bertujuan untuk membuat program berupa peringkasan otomatis menggunakan *Recurrent Neural Networks (RNN) Encoder-Decoder* yang bersifat abstraktif. *RNN Encoder-Decoder* dipilih sebagai metode untuk peringkasan karena metode ini sudah banyak berhasil untuk membuat teks, mesin penerjemah, dan pengenalan suara. Contoh peringkasan yang menggunakan metode *RNN Encoder-Decoder* adalah peringkasan yang dibuat oleh Ramesh Nallapati et al [3].

RNN Encoder-Decoder adalah metode yang dikembangkan oleh Cho et al. Metode ini dapat mengubah teks masukan menjadi teks lainnya. Rangkaian x dimasukkan ke dalam *Encoder* sehingga menghasilkan sebuah vektor c yang berisi nilai real dengan ukuran vektor yang tetap. Vektor c adalah vektor yang merepresentasi teks masukan. Kemudian vektor c dimasukkan ke dalam *decoder* sehingga menghasilkan keluaran y [4]. Peringkasan dengan metode ini memerlukan metode untuk membuat representasi kata. Metode yang dipilih adalah *Word2vec*. Metode ini dipilih karena *Word2vec* dapat membuat representasi kata berdasarkan konteks-konteks di sekitarnya dengan kualitas tinggi dan efisien [5]. Penelitian ini terfokus untuk menghitung akurasi dari metode *RNN Encoder-Decoder* dan mencari dan mendapatkan nilai optimal dari variabel-variabel yang terdapat pada *RNN Encoder-Decoder* dan *Word2vec*. Variabel-variabel yang diteliti yaitu jumlah *hidden state Word2vec* dan jumlah *hidden state RNN Encoder-Decoder*.

2. METODOLOGI

2.1 Ringkasan

Ringkasan didefinisikan sebagai penyajian bacaan dalam bentuk singkat. Ringkasan dibuat dengan mempertahankan urutan isi dan sudut pandang. Ringkasan bisa juga dibuat dengan memendekkan bacaan dengan mengambil inti sari bacaan itu tanpa mengubah struktur wacana [6].

2.2 RNN

Recurrent Neural Networks (RNN) adalah bagian dari jaringan saraf tiruan yang dapat memproses data sekuensial. RNN dapat memproses rangkaian yang berupa x_1, \dots, x_T . RNN mengeluarkan rangkaian keluaran dengan jumlah elemen yang sama dengan jumlah elemen pada sekuens masukan [7].

RNN dalam penelitian ini menggunakan keluaran diskrit untuk memprediksi kata atau karakter. Cara paling alami untuk merepresentasi variabel diskrit adalah memperhatikan keluaran o dengan memberikan log probabilitas yang belum normal untuk setiap nilai yang mungkin dalam variabel diskrit. Operasi *softmax* dilakukan sebagai tahap pasca-proses untuk mendapatkan vektor \hat{y} yang berupa probabilitas keluaran yang telah dinormalisasi. Propagasi maju dimulai dengan memberikan nilai pada *state* h_0 . *Softmax* merupakan fungsi yang sering digunakan sebagai keluaran untuk merepresentasi distribusi probabilitas dari n kelas yang

berbeda. *Softmax* dapat digunakan satu dari beberapa n pilihan berbeda untuk beberapa variabel internal. n digunakan sebagai jumlah seluruh kata dalam penelitian ini.[7]

Arsitektur yang umum digunakan adalah *Elman Network* atau biasa disebut *simple RNN*. Untuk tiap-tiap waktu dari $t = 1$ sampai $t = T$, persamaan berikut diterapkan:

$$h_t = \tanh(W \cdot h_{t-1} + U \cdot x_t) \quad (1)$$

$$o_t = V \cdot h_t \quad (2)$$

$$\hat{y}_t = \text{softmax}(o_t) \quad (3)$$

$$\text{softmax}(z)_i = \frac{\exp(z_i)}{\sum_{j=1}^n \exp(z_j)} \quad (4)$$

Dengan W , U , dan V adalah parameter bobot RNN, h_t adalah *hidden state*, o_t adalah *output RNN*, dan \hat{y}_t adalah probabilitas kata berdasarkan o_t .

2.3 Encoder-Decoder

Encoder-Decoder adalah jaringan saraf buatan yang terdiri dari dua buah RNN. *Encoder-Decoder* bisa belajar untuk mengodekan (*encode*) rangkaian ke representasi vektor dengan panjang yang ditentukan. Vektor tersebut diterjemahkan (*decode*) kembali ke rangkaian yang lain sehingga panjang dari rangkaian dan masukkan dapat berbeda. Hal ini memungkinkan untuk membuat mesin penerjemah dan mesin peringkas otomatis [4].

Decoder merupakan RNN yang dilatih untuk membuat rangkaian keluaran dengan memprediksi kata y_t berikutnya berdasarkan *hidden state* h_t . Namun, RNN ini tidak seperti RNN biasanya. Variabel y_t dan h_t juga dikondisikan oleh y_{t-1} dan ringkasan c dari rangkaian masukan. Oleh karena itu, *hidden state* dari *decoder* pada waktu t adalah:

$$h_t = f(h_{t-1}, y_{t-1}, c) \quad (5)$$

dengan f adalah fungsi aktivasi.

2.4 Word2vec

Word2vec adalah metode untuk membuat *word embedding* dengan berdasarkan pada RNN *Language Model* (RNNLM). Metode ini mampu mempelajari vektor representasi kata dengan kualitas tinggi dari banyak data teks yang tidak terstruktur [5].

Vektor masukan w_t merepresentasi kata masukan pada waktu t yang dikodekan dengan *1-of-N*. Lapisan keluaran y_t memproduksi distribusi probabilitas terhadap kata-kata. Lapisan *hidden* s_t mempertahankan representasi dari sejarah kalimat. Perbedaan RNN dengan RNNLM terdapat pada perhitungan lapisan *hidden*. Nilai-nilai pada lapisan *hidden* dihitung sebagai berikut:

$$s(t) = f(U \cdot w(t) + W \cdot s(t-1)) \quad (6)$$

$$y(t) = \text{softmax}(V \cdot s(t)) \quad (7)$$

$$f(z) = \frac{1}{1 + e^{-z}} \quad (8)$$

2.5 Long-Short Term Memory

Long-Short Term Memory (LSTM) adalah arsitektur RNN yang umum digunakan. LSTM dapat mengakumulasi informasi dalam waktu yang panjang. LSTM juga dapat melupak *state* lama yang telah digunakan secara otomatis. Bobot pada LSTM dikondisikan berdasarkan

konteks dan dikendalikan oleh *hidden unit* lainnya. LSTM ditemukan sukses pada beberapa aplikasi seperti *speech recognition*, *handwriting generation* dan *machine translation* [7].

LSTM memiliki *recurrence* di dalam selnya sebagai tambahan dari *recurrence* dari RNN. Tiap-tiap sel memiliki *input* dan *output* yang sama seperti RNN pada umumnya tetapi memiliki parameter lebih banyak dan sistem gerbang unit yang mengontrol aliran informasi. Komponen terpenting dari LSTM adalah *state unit* yang memiliki *linear self-loop* yang dikendalikan oleh *forget gate unit*. Rumus yang digunakan adalah sebagai berikut:

$$hf_t = \sigma(W_f \cdot [h_{t-1}; x_t] + b_f) \quad (9)$$

$$hi_t = \sigma(W_i \cdot [h_{t-1}; x_t] + b_i) \quad (10)$$

$$ho_t = \sigma(W_o \cdot [h_{t-1}; x_t] + b_o) \quad (11)$$

$$hc_t = \tanh(W_c \cdot [h_{t-1}; x_t] + b_c) \quad (12)$$

$$c_t = hf_t * c_{t-1} + hi_t * hc_t \quad (13)$$

$$h_t = ho_t * \tanh(c_t) \quad (14)$$

Dengan hf_t adalah *forget gate unit*, hi_t adalah *external gate unit*, ho_t adalah *output gate*, hc_t adalah *hidden state*, c_t adalah *cell state*, dan h_t adalah *output*.

2.6 Attention-Mechanism

Mekanisme *attention* digunakan setelah *layer* LSTM untuk meningkatkan kualitas hasil ringkasan dengan memilih bagian-bagian dari teks asli. *Attention* adalah metode yang dapat membuat model mempelajari keselarasan di antara modalitas yang berbeda misalnya di antara *speech frames* dengan teks. *Neural Machine Translation* dapat mengaplikasikan mekanisme *attention* untuk menerjemahkan dan menyelaraskan kata-kata secara bersama-sama [8].

Mekanisme *attention* menerima *input* berupa *hidden state* h_t dari layer sebelumnya. Tujuannya adalah untuk mendapatkan konteks vektor c_t yang menangkap informasi relevan dari teks asli untuk membantu memprediksi target kata y_t saat ini. Rumus yang digunakan adalah sebagai berikut:

$$a_t(s) = \frac{\exp(\text{score}(h_t, \bar{h}_s))}{\sum_{s'} \exp(\text{score}(h_t, \bar{h}_{s'}))} \quad (15)$$

$$\text{score}(h_t, \bar{h}_s) = h_t^T \bar{h}_s \quad (16)$$

$$c_t = \sum_s a_t(s) \bar{h}_s \quad (17)$$

$$\tilde{h}_t = \tanh(W_c \cdot [c_t; h_t]) \quad (18)$$

$$O_t = W_s \tilde{h}_t \quad (19)$$

$$y_t = \text{softmax}(O_t) \quad (20)$$

Dengan h_t adalah *hidden state*, \bar{h}_s adalah *hidden state* dari teks asli, $a_t(s)$ adalah *alignment vector* skor antara \bar{h}_s dengan h_t , \tilde{h}_t adalah *hidden state attention*, O_t adalah *output*, dan y_t adalah probabilitas kata berdasarkan O_t .

3. PENGUJIAN

Data yang diperlukan dalam penelitian ini adalah koleksi berita dan koleksi ringkasan. Berita dikumpulkan dari situs www.kompas.com sejumlah 180 berita atau disebut dengan koleksi berita kompas dan www.detik.com sejumlah 6273 berita atau disebut dengan koleksi

berita detik. Total berita yang dikumpulkan adalah 6453 berita. Kedua koleksi berita tersebut dikumpulkan dengan menggunakan program *scraper* yang dapat mengumpulkan berita tanpa perlu menerima seluruh *tag* html-nya.

Pembuatan Word2vec dalam penelitian ini menggunakan data latih berupa koleksi berita detik dan koleksi berita kompas. Hasil dari pelatihan Word2vec adalah 61497 token untuk koleksi berita detik dan 6197 token untuk koleksi berita kompas dengan vektor *word embedding*-nya. *Word embedding* ini digunakan sebagai bobot sebuah kata di dalam koleksi berita.

Koleksi pertanyaan yang digunakan untuk Q&A berjumlah 250 pertanyaan. Pertanyaan-pertanyaan ini terdiri dari lima pertanyaan untuk lima puluh berita data uji. Data uji ini terdiri dari 70% data latih dan 30% data dari luar data uji.

Pengujian sistem dilakukan dengan menggunakan data kompas dan data detik. Ringkasan asli dari data latih dibandingkan dengan ringkasan sistem. Jika semua token dari ringkasan sistem sama dengan ringkasan asli maka jumlah ringkasan benar bertambah satu. Akurasi dari model adalah jumlah ringkasan benar dibagi dengan jumlah seluruh ringkasan dikalikan dengan 100.

Tabel 1 Hasil pelatihan dan pengujian data kompas

nama model	<i>Epoch</i>	jumlah <i>hidden state word2vec</i>	Jumlah <i>hidden state RNN</i>	<i>loss</i>	Akurasi
model 1	5000	20	20	0.0871	0%
model 2	5000	20	50	0.0586	39.0109%
model 3	5000	50	100	0.006	34.6153%

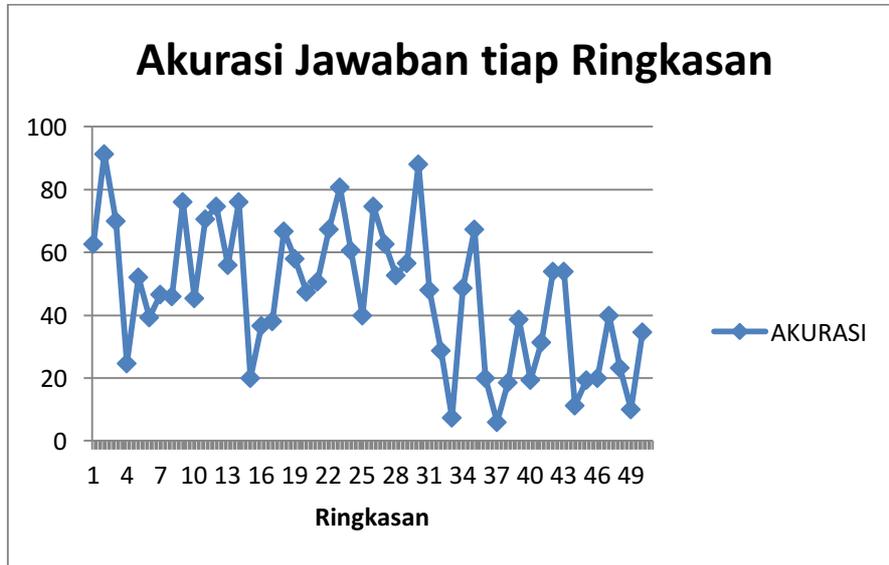
Tabel 1 menunjukkan hasil pelatihan dan pengujian dengan menggunakan data kompas. Pelatihan menghasilkan tiga model yaitu model 1 dengan jumlah word2vec sebanyak 20 dan jumlah *hidden state* sebanyak 20, model 2 dengan jumlah word2vec sebanyak 20 dan jumlah *hidden state* sebanyak 50, dan model 3 dengan jumlah word2vec sebanyak 50 dan jumlah *hidden state* sebanyak 100. Model 1 memiliki akurasi sebesar 0%, model 2 memiliki akurasi sebesar 39.0109%, dan model 3 memiliki akurasi sebesar 34.6153%. Tiap-tiap model memiliki nilai *loss* di bawah 1. Namun, akurasi tertinggi yang didapatkan dengan data kompas adalah 39.0109%. Oleh karena itu, pelatihan selanjutnya dilakukan dengan menggunakan data detik.

Tabel 2 Hasil pelatihan dan pengujian data detik

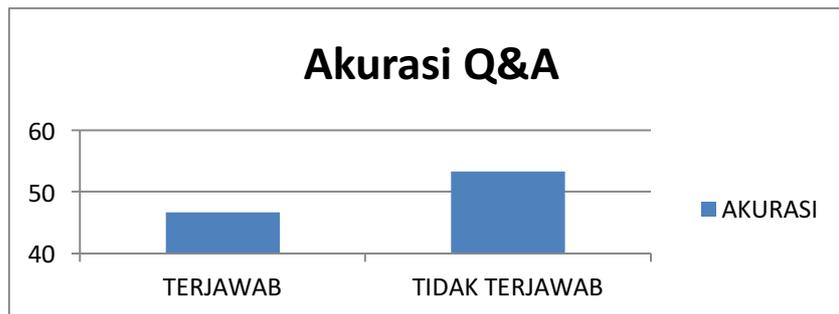
nama model	<i>Epoch</i>	Jumlah <i>hidden state word2vec</i>	Jumlah <i>hidden state RNN</i>	<i>loss</i>	akurasi
model 4	609	100	100	0.0151	55.4481%
model 5	609	200	256	0.0006654	99.8810%
model 6	609		512	0.02788	50.2898%

Tabel 2 menunjukkan hasil pelatihan dan pengujian dengan menggunakan data detik. Pelatihan menghasilkan tiga model yaitu model 4 dengan jumlah word2vec sebanyak 100 dan jumlah *hidden state* sebanyak 100, model 5 dengan jumlah word2vec sebanyak 200 dan jumlah *hidden state* sebanyak 256, dan model 6 dengan jumlah word2vec sebanyak 200 dan jumlah *hidden state* sebanyak 256. model 4 memiliki akurasi sebesar 55.4481%, model 5 memiliki akurasi sebesar 99.8810%, dan model 6 memiliki akurasi sebesar 50.2898%. Tiap-tiap model memiliki nilai *loss* di bawah 1. Akurasi tertinggi yang didapatkan dengan data kompas adalah 99.8810%.

Pengujian Q&A Evaluation menghitung akurasi dari model terbaik yang dipilih pada pengujian sistem dengan 250 pertanyaan dari data uji yang berjumlah 50 data dengan 5 pertanyaan untuk tiap-tiap data. Evaluasi ini dilakukan terhadap 30 responden yang terdiri dari mahasiswa. Hasil dari Q&A Evaluation dapat dilihat pada tabel berikut:



Gambar 1 Akurasi Jawaban Tiap Ringkasan dalam persen (%)



Gambar 2 Hasil Evaluasi Q&A Evaluation dalam persen (%)

Berdasarkan Gambar 1, tiap-tiap dokumen memiliki akurasi yang berbeda-beda. Akurasi ringkasan cenderung berada di antara 20% sampai 60%. Setelah dirata-ratakan, sistem peringkasan otomatis abstraktif dapat menghasilkan ringkasan dengan akurasi 46,65%.

4. Pembahasan

Pembahasan dibagi menjadi dua yaitu pembahasan pengujian sistem dan pembahasan pengujian Q&A Evaluation. Pembahasan dilakukan dengan membandingkan berita dengan ringkasan sistem. Ringkasan sistem dibuat dari model-model yang telah dilatih.

Hasil ringkasan tiap-tiap model diperlukan untuk melihat dan meneliti keberhasilan dan kesalahan sistem. Hasil ringkasan sistem juga menggambarkan seberapa baik sistem dapat mengenali berita dan membuat kata-kata yang sesuai dengan tema berita dan dapat dimengerti oleh manusia. Oleh karena itu, contoh hasil ringkasan tiap-tiap model dapat dilihat sebagai berikut:

Hasil peringkasan pada salah satu data latih dengan menggunakan model 1 adalah sebagai berikut:

Wakil Kepala Polda Metro Jaya Brigjen Pol Suntana menyatakan, Polda Metro Jaya masih menunggu Gubernur DKI Jakarta Djarot Saiful Hidayat meminta aturan kepemilikan garasi sebagai syarat mendapatkan surat tanda nomor kendaraan (STNK) diterapkan. Aturan tentang kewajiban pemilik kendaraan bermotor memiliki atau menguasai garasi tertuang dalam Peraturan Daerah (Perda) Nomor 5 Tahun 2014 tentang Transportasi. " Kami tinggal menunggu perintah dari Pak Gubernur," ujar Suntana, di Balai Kota DKI Jakarta, Jalan Medan Merdeka Selatan, Rabu (13 / 9 / 2017). Suntana belum mengetahui kapan aturan tersebut mulai diterapkan bagi warga yang ingin membeli kendaraan bermotor. Sebab, Pemprov DKI Jakarta saat ini masih menyosialisasikan aturan tersebut. " Saya belum tahu, nanti Pak Gubernur. Nanti kami akan diskusikan secara bertahap ya, ini kan lagi sosialisasi," kata Suntana. Meski begitu, Suntana menyebut Polda Metro Jaya akan mendukung dan melaksanakan perda tersebut. Sementara itu, Djarot juga belum menyebut kapan realisasi aturan untuk penerbitan kendaraan bermotor yang baru ditegakkan. Dia hanya menekankan kepemilikan garasi harus dibuktikan dengan surat dari kelurahan. " Kalau penerbitan STNK baru, mobil baru kan, ya itu ada jaminan, sepengetahuan RT, RW, sampai kelurahan. Baru setelah ada jaminannya, baru kami sampaikan ke samsat untuk diproses STNK," ujar Djarot .

Gambar 3 Berita 1 berjudul “Polda Metro Tunggu Djarot Perintahkan Penerapan Kepemilikan Garasi sebagai Syarat Mendapat STNK” (Sumber: Nursita Sari, Polda Metro Tunggu Djarot Perintahkan Penerapan Kepemilikan Garasi sebagai Syarat Mendapat STNK, <http://megapolitan.kompas.com/read/2017/09/13/13511081/polda-metro-tunggu-djarot-perintahkan-penerapan-kepemilikan-garasi>, 13 September 2017)

Ringkasan:

Polisi telah menangkap tiga pembunuh Husni Zarkasih dan istrinya Zakiah Husni. Ketiganya ditangkap di kawasan Grobogan, Jawa Tengah. Korban merupakan pengusaha garmen di Pekalongan, Jawa Tengah. Korban merupakan pengusaha garmen di Pekalongan, Jawa Tengah. Berdasarkan informasi yang dihimpun, seorang pelaku merupakan mantan sopir yang dipecah korban karena suatu masalah.

Gambar 4 Hasil peringkasan berita 1 dengan model 1

Hasil peringkasan pada salah satu data latih dengan menggunakan model 2 adalah sebagai berikut:

Berita:

Ketua DPRD DKI Jakarta UNK UNK UNK meminta Dinas Kehutanan, Pertamanan, dan Pemakaman DKI Jakarta menata pedagang bunga yang berjualan di tempat PAD umum (UNK) di Jakarta. Penataan UNK agar seluruh area UNK lebih rapi. " Pedagang bunga itu UNK UNK lokasi yang bagus seperti pasar UNK di Blok M, bagus," ujar UNK, dalam rapat Badan Anggaran (UNK) bersama UNK di Gedung DPRD DKI Jakarta, Jumat (25/8/2017). Kepala Dinas Kehutanan, Pertamanan, dan Pemakaman DKI Jakarta UNK UNK menyebut penataan pedagang bunga sudah UNK di beberapa UNK bekerja sama dengan Dinas Koperasi, UMKM, dan Perdagangan. UNK mengungkapkan, penataan akan terus dilakukan karena masih banyak lokasi berjualan pedagang bunga yang UNK. " Memang sudah kami UNK dengan UMKM supaya jadi binaan UMKM untuk disediakan UNK, UNK rapi, sehingga di UNK atau halaman UNK itu rapi," kata UNK, se usai rapat. Penataan para pedagang bunga di UNK, kata UNK, akan menggunakan dana yang berasal dari program UNK UNK UNK (UNK) UNK .

Gambar 5 Berita 2 berjudul “Pimpinan DPRD DKI Minta Pedagang Bunga di Seluruh TPU Ditata” (Sumber: Nursita Sari, Pimpinan DPRD DKI Minta Pedagang Bunga di Seluruh TPU Ditata, <http://megapolitan.kompas.com/read/2017/08/25/20471291/pimpinan-dprd-dki-minta-pedagang-bunga-di-seluruh-tpu-ditata>, 13 September 2017)

Ringkasan:

Dinas Perumahan DKI Jakarta Agustino Darmawan akan menghapus denda untuk warga yang menunggak biaya sewa rumah susun. Polisi telah warga rusun .

Gambar 6 Hasil peringkasan berita 2 dengan model 2

Polisi memutuskan untuk membebaskan seorang karyawan Diskotek Diamond yang sebelumnya ditangkap. UNK tersebut diduga menyediakan sabu untuk politisi Partai Golkar Indra J Piliang dan dua rekannya. Kabid Humas Polda Metro Jaya Kombes Argo Yuwono mengatakan, karyawan tersebut UNK setelah melewati UNK pemeriksaan. " Enggak (UNK). Setelah UNK jam (diperiksa) kita UNK," ujar Argo di kawasan Monas, Jakarta Pusat, Sabtu (16/9/2017). Argo menambahkan, karyawan tersebut UNK karena penyidik belum menemukan bukti kuat bahwa dia merupakan UNK sabu untuk Indra. Menurut Argo, dalam pemeriksaan kemarin karyawan Diskotek Diamond itu tak mengakui perbuatannya. " Kita lakukan UNK karena tidak semua orang langsung mengatakan yang UNK . Barang UNK saja belum ada, ya kita tunggu saja. Masih kita UNK semua," kata Argo. Indra ditangkap bersama dua rekannya, Romi Fernando dan M Ismail Jamani, pada Rabu (13/9/2017), sekitar pukul 19. 30 WIB, di Diskotek Diamond. Dari penangkapan Indra dan kedua rekannya itu, polisi mengamankan barang bukti berupa satu set alat isap sabu dan UNK bekas pakai, satu plastik bekas pakai dan sebuah korek api. Setelah dilakukan tes urine, Indra dinyatakan positif mengonsumsi sabu. Indra sudah ditetapkan menjadi tersangka dan akan menjalani rehabilitasi.

Gambar 7 Berita 3 berjudul "Polisi Bebaskan Karyawan Diskotek Diamond Setelah Diperiksa"

(Sumber: Akhdi Martin Pratama, Polisi Bebaskan Karyawan Diskotek Diamond Setelah Diperiksa, <http://megapolitan.kompas.com/read/2017/09/16/16201771/polisi-bebaskan-karyawan-diskotek-diamond-setelah-diperiksa>, 13 September 2017)

Ringkasan:

Politisi Partai Golkar, mengatakan, kondisi jalan layang Koridor 13 Transjakarta. Antrean calon penumpang terjadi. Polisi ditangkap yang ada saat itu turut rumah sakit rumah sakit yang sudah titik yang membutuhkan.

Gambar 8 Hasil ringkasan berita 3 dengan model 3

Berdasarkan Gambar 4, Gambar 6, dan Gambar 8, hasil peringkasan memiliki makna yang jauh dari berita aslinya walaupun memiliki nilai *loss* di bawah 1. Jumlah *word2vec* dan jumlah *hidden state* yang terlalu sedikit pada model 1 memiliki akurasi yang sangat kecil dan hasil ringkasan dengan topik yang sangat berbeda dari berita aslinya. Jumlah *word2vec* dan jumlah *hidden state* pada model 2 dan model 3 tidak memiliki banyak pengaruh terhadap akurasi dan hasil ringkasan yang dibuat masih memiliki topik yang berbeda dari berita aslinya. Namun, model 2 dan model 3 sudah dapat mengenali subyek dari berita. Pada Gambar 6, ringkasan membahas tentang denda warga di rumah susun sedangkan berita pada Gambar 5 membahas tentang pedagang bunga. Model 2 dapat menghasilkan kata "Dinas Perumahan" walaupun subjek pada Gambar 5 adalah "Dinas Kehutanan". Pada Gambar 7 dan Gambar 8, model 3 dapat menghasilkan kata "Politisi Partai Golkar" walapun topik pada Gambar 7 membahas tentang politisi partai golkar yang terlibat narkoba dan Gambar 8 membahas tentang transjakarta.

Selain itu, masih banyak kata yang tidak dikenal dalam peringkasan. Kata-kata yang tidak dikenal ini ditandai dengan simbol UNK. Hal ini disebabkan oleh jumlah kumpulan kata yang dimiliki yaitu sebesar 6149 kata masih kurang banyak.

Data detik digunakan dalam pelatihan karena memiliki jumlah data yang banyak. Data detik bisa menghasilkan 61497 token yang terdiri dari kata dan simbol sehingga bisa mengurangi kata-kata yang tidak dikenal. Namun, jumlah token yang banyak ini mengakibatkan program berjalan lebih lama dibandingkan dengan menggunakan data kompas.

Berdasarkan tabel 5, terlihat bahwa untuk mendapatkan akurasi yang tinggi terhadap data latih diperlukan jumlah *word2vec* dan jumlah *hidden state* yang tepat. Jumlah yang terlalu tinggi dan terlalu rendah bisa memiliki akurasi yang serupa.

Hasil ringkasan dari model 4, model 5, dan model 6 pada tabel 5 dapat dilihat sebagai berikut:

Berita:

Gubernur DKI Jakarta Djarot Saiful Hidayat meresmikan revitalisasi Kota Tua tahap I. Djarot mengatakan revitalisasi tersebut merupakan kontribusi dari gagasan gubernur sebelumnya, yakni Joko Widodo dan Basuki Tjahaja Purnama (Ahok). "Proses hari ini tidak bisa dilepaskan dari gagasan sejak awal pada 2012, Pak Jokowi, kemudian dikonkretkan Pak Ahok sampai 2017, dan saya yang menutup saat ini," kata Djarot saat di kawasan Kota Tua, Jakarta Barat, Kamis (5/10/2017). Revitalisasi yang dilakukan meliputi normalisasi dan pembenahan trotoar di Kali Besar, kemudian membenahi gedung Kerta Niaga menjadi lokasi berjualan, serta memindahkan pedagang kaki lima ke lokasi baru di Jalan Cengkeh. Kota Tua juga dipercantik dengan pemasangan lima patung baru untuk mempercantik pemandangan. Gubernur DKI Jakarta Djarot Saiful Hidayat meresmikan revitalisasi Kota Tua tahap I. (Fida detikcom) Djarot, yang hadir bersama jajaran satuan kerja perangkat daerah (SKPD) terkait beserta Wali Kota Jakarta Barat Anas Effendi, berkeliling mengecek hasil renovasi dan pembaruan Kota Tua. Djarot melepas 5.000 benih ikan ke sungai sekitar Kali Baru. Djarot mengatakan kawasan Kali Baru dipercantik dengan keberadaan Jembatan Budaya yang dapat menjadi pusat warga berkeaktivitas. Keberadaan jembatan itu diharapkan menjadi sarana pemersatu warga Jakarta. "Jembatan Budaya ini berarti Jembatan Budaya Indonesia yang beragam beraneka warna dan bersatu," sebutnya. Djarot juga ikut mengecek lokasi binaan PKL yang berada di Jalan Cengkeh. Lokasi tersebut berjarak sekitar 200 meter dari Museum Fatahillah. Terdapat 456 PKL yang di antaranya diisi 128 pedagang kuliner. Djarot berharap Kota Tua dapat menjadi pusat wisata kuliner dan sejarah. "Ini wisata yang sehat kuliner karena berjalan kaki dapat semakin mengakrabkan warga," papar Djarot. Djarot menyebut peresmian kali ini merupakan kejutan darinya menjelang akhir masa jabatan. Pihaknya berharap pemerintahan Anies Baswedan-Sandiaga Uno dapat melanjutkan program revitalisasi tahap II dan III sesuai yang direncanakan. "Kami berharap tahap selanjutnya dapat dilanjutkan pemerintahan yang akan dilantik," jelasnya.

Gambar 9 Berita 4 berjudul "Revitalisasi Kota Tua Tahap I, Djarot: Ini Gagasan Jokowi-Ahok" (Sumber: Muhammad Fida UI Haq, Revitalisasi Kota Tua Tahap I, Djarot: Ini Gagasan Jokowi-Ahok, <https://news.detik.com/berita/d-3672340/revitalisasi-kota-tua-tahap-i-djarot-ini-gagasan-jokowi-ahok>, 5 November 2017)

Ringkasan:

Plt DKI Jakarta Sandiaga Saiful Hidayat meresmikan revitalisasi Kota Tua tahap I rumah berisi uang Rp (2) besok besok .

Gambar 10 Hasil ringkasan berita 4 dengan model 4

Berita:

Gubernur DKI Jakarta Djarot Saiful Hidayat mengatakan, Pemprov DKI Jakarta akan menandatangani perjanjian kerja sama dengan Polda Metro Jaya terkait aturan kepemilikan garasi sebagai syarat untuk menerbitkan surat tanda nomor kendaraan (STNK). Aturan tentang kewajiban pemilik kendaraan bermotor memiliki atau menguasai garasi tertuang dalam Peraturan Daerah (Perda) Nomor 5 Tahun 2014 tentang Transportasi. "Besok kami juga akan kerja sama dengan Polda, untuk penerbitan STNK ini butuh ada perjanjian kerja sama dengan Polda. Mereka yang betul-betul punya garasi, ada jaminan punya garasi, itu bisa ada STNK-nya," ujar Djarot di Balai Kota DKI Jakarta, Jalan Medan Merdeka Selatan. Djarot menyebut Dinas Perhubungan DKI Jakarta sudah gencar menyosialisasikan aturan tersebut. Aturan kepemilikan garasi itu dilakukan untuk mendorong ketertiban di Jakarta. "Untuk pemilik garasi ini kami sudah mulai sosialisasikan dan saya lihat Dinas Perhubungan sudah mulai ada beberapa yang ditangkap," kata Djarot. Perda Nomor 5 Tahun 2014 menyebutkan pemilik kendaraan bermotor wajib memiliki atau menguasai garasi dan tidak boleh memarkir kendaraan di jalan. Warga atau badan usaha yang akan membeli kendaraan bermotor wajib memiliki atau menguasai garasi yang dibuktikan dengan surat bukti kepemilikan garasi dari kelurahan setempat. Surat bukti itu menjadi syarat penerbitan STNK.

Gambar 11 Berita 6 berjudul "Aturan Kepemilikan Garasi sebagai Syarat Penerbitan STNK Akan Diterapkan" (Sumber: Nursita Sari, Aturan Kepemilikan Garasi sebagai Syarat Penerbitan STNK Akan Diterapkan, <http://megapolitan.kompas.com/read/2017/09/11/15465361/aturan-kepemilikan-garasi-sebagai-syarat-penerbitan-stnk-akan-diterapkan>, 13 September 2017)

Ringkasan:

Plt Gubernur DKI Jakarta Sandiaga Uno terkait dengan narkoba pornografi di Gedung .

Gambar 12 Hasil ringkasan berita 6 dengan model 6

Gubernur DKI Jakarta Djarot Saiful Hidayat meresmikan revitalisasi tahap pertama Kawasan Kota Tua, Tamansari, Jakarta Barat, Kamis (5/10/2017) malam. Salah satu revitalisasi tahap pertama ini yakni pembangunan Jembatan Budaya di atas Kali Krukut yang menghubungkan Jalan Kali Besar Timur dan Jalan Kali Besar Barat. Peresmian itu ditandai dengan memecahkan dua buah kendi. Jembatan dan kendi dalam peresmian Kawasan Kota Tua itu sarat makna akan budaya Indonesia. "Jembatan budaya ini berarti jembatan budaya Indonesia yang beragam, beraneka warna, dan bersatu," ujar Djarot di Jembatan Budaya. Tak hanya itu, Djarot juga menyebutkan makna khusus dari kendi yang dia pecahkan. Kendi itu dia namakan kendi nusantara. "Kendi ini sebagai bagian dari nusantara, dari Sabang sampai Merauke. Mari kita nikmati sama-sama," kata Djarot sambil memecahkan kedua kendi itu. Setelah memecahkan kendi, Djarot bersama pejabat Pemprov DKI Jakarta melepas sekitar 5.000 benih ikan ke Kali Krukut sebagai tanda peresmian tersebut. Kepala Dinas Pariwisata dan Kebudayaan DKI Jakarta Tinia Budiati menyampaikan, pembangunan Jembatan Budaya merupakan bagian dari revitalisasi Kawasan Kali Besar yang ada di samping Kawasan Taman Fatahillah. "Untuk tahap pertama telah dilaksanakan penataan Kali Besar dari Jalan Bank ke Jalan Kopi dengan memasang patung karya Ibu Dolorosa (Seniman Dolorosa Sinaga)," tutur Tinia. Revitalisasi Kawasan Kali Besar dilengkapi dengan trotoar selebar sembilan meter di sisi kiri dan kanan Kali Krukut. Revitalisasi Kali Besar dibangun menggunakan dana kewajiban kompensasi atas pelampauan nilai koefisien lantai bangunan (KLB) PT Sampoerna Land. Selain Jembatan Budaya, Djarot juga meresmikan Gedung Kerta Niaga, Museum Sejarah Jakarta, dan Lokasi Binaan Taman Kota Intan di Jalan Cengkeh yang merupakan bagian dari revitalisasi tahap pertama Kawasan Kota Tua.

Gambar 13 Berita 5 berjudul "Jembatan dan Kendi yang Sarat Makna Budaya dalam Peresmian Kota Tua" (Sumber: Nursita Sari, Jembatan dan Kendi yang Sarat Makna Budaya dalam Peresmian Kota Tua, <http://megapolitan.kompas.com/read/2017/10/05/22151681/jembatan-dan-kendi-yang-sarat-makna-budaya-dalam-peresmian-kota-tua>, 12 November 2017)

Ringkasan:

Gubernur DKI Jakarta Djarot Saiful Hidayat meresmikan revitalisasi Kota Tua tahap I .

Gambar 14 Hasil ringkasan berita 5 dengan model 5

Berdasarkan hasil ringkasan pada Gambar 10, Gambar 12, dan Gambar 14, model yang memiliki topik ringkasan dan makna sesuai dengan berita aslinya adalah model 5 dengan topik revitalisasi kota tua tahap 1. Pada Gambar 10, ringkasan memiliki topik yang sama dengan Gambar 9 tetapi subjeknya berbeda karena "Djarot Saiful Hidayat" menjadi "Sandiaga Saiful Hidayat". Pada Gambar 12, ringkasan memiliki topik yang berbeda dengan Gambar 11. Gambar 12 membahas tentang narkoba dan pornografi sedangkan Gambar 11 membahas tentang peraturan STNK. Meskipun demikian, Gambar 12 memiliki subjek yang mirip dengan Gambar 11 yaitu "Djarot Saifu Hidayat" menjadi "Sandiaga Uno".

Berita-berita yang diolah oleh program sudah tidak sering muncul token UNK. Hal ini terjadi karena jumlah token yang cukup banyak yaitu berjumlah 61497 token. Oleh karena itu, program bisa mengenali berita dengan lebih baik.

Pengujian Q&A Evaluation menunjukkan akurasi sistem sebesar 46,65%. Akurasi ini kecil karena ada beberapa ringkasan yang tidak bisa menjawab pertanyaan-pertanyaan yang diberikan. Ringkasan-ringkasan memiliki beberapa kata yang berhubungan dengan berita aslinya. Namun, kalimat yang dibuat tidak sesuai dengan berita aslinya. Kemudian, banyak ringkasan yang hanya bisa menjawab dua sampai tiga pertanyaan saja.

5. KESIMPULAN

Berikut beberapa hal yang dapat disimpulkan dari penelitian sistem peringkat otomatis abstraktif dengan RNN yang telah dilakukan, antara lain:

1. Pengujian menunjukkan bahwa model terbaik diperoleh dengan jumlah *word2vec* sebanyak 200 dan jumlah *hidden state* sebanyak 256 dengan *loss* sebesar 0.0006654 dan akurasi sebesar 99.8810% untuk pengujian sistem dan akurasi sebesar 46,65% untuk pengujian Q&A Evaluation.
2. Sistem dapat menghasilkan ringkasan yang baik jika topik dari berita yang diringkas mirip dengan topik berita yang berada di dalam data latih walaupun kata-kata pada berita berbeda dengan data latih.
3. Kesalahan yang dibuat sistem berupa ringkasan dengan kata-kata yang berhubungan subjek atau lokasi dari berita.

DAFTAR PUSTAKA

- [1] Kusmayadi, Ismail, 2008, "Think Smart Bahasa Indonesia", Grafindo Media Pratama, Bandung.
- [2] Ridok, Achmad, April 2014, "Peringkasan Dokumen Bahasa Indonesia Berbasis Non-Negative Matrix Factorization", Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK), Vol.1, Nomor 1, Jakarta.
- [3] Nallapati, Ramesh; Zhou, Bowen; Cicero, Santos dos; Gulçehre, Çağla and Xiang, Bing, Agustus 2016, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond", Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), Berlin-German.
- [4] Cho, Kyunghyun; Merriënboer, Bart van; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger and Bengio, Yoshua, Oktober 2014, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha-Qatar.
- [5] Mikolov, Tomas, Desember 2013, "Distributed Representations of Words and Phrases and their Compositionality", NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems, Stateline-USA.
- [6] Zaman B., Winarko E., Juli 2011 "Analisis Fitur Kalimat untuk Peringkat Teks Otomatis pada Bahasa Indonesia", Indonesian Journal of Computing and Cybernetics Systems, vol. 5, nomor 2, Yogyakarta.
- [7] Goodfellow Ian, Bengio Yoshua, Courville Aaron, 2006, Deep Learning, MIT Press, Massachusetts.
- [8] Luong Minh-Thang, Pham Hieu, Manning Christopher D., September 2015, "Effective Approaches to Attention-based Neural Machine Translation", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon-Portugal