LAPORAN TUGAS BESAR
PEMROSESAN BAHASA ALAMI

Klasifikasi Teks Ulasan Menjadi Bentuk Rating Menggunakan Metode *Long Short-term Memory*

Disusun Oleh:

Aditya Alif Nugraha	1301154183
Nadine Azhalia Purbani	1301154519

ICM-GAB01

Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2018

I. Abstrak

Produk kecantikan saat ini sangat sering digunakan dikalangan wanita. Adanya suatu produk tidak terlepas dari ulasan atau komentar yang diberikan oleh penggunanya. Ulasan suatu produk merupakan pengalaman pribadi yang dialami oleh konsumen setelah menggunakan produk. Ulasan tersebut dapat diterjemahkan menjadi bentuk rating yang ditentukan sendiri oleh konsumen. Rating yang diberikan oleh konsumen bersifat personal, artinya konsumen memiliki pandangan tersendiri dalam menentukan rating. Maka dari itu dibutuhkan sistem yang dapat menerjemahkan ulasan menjadi bentuk rating agar rating yang didapat bersifat general. Untuk membangun sistem digunakan metode *Long Short-Term Memory* (LSTM). Ulasan yang diberikan akan dikelompokan menjadi lima kelas rating, semakin kecil nilai rating menandakan konsumen sangat tidak puas dengan produk. Berlaku pula untuk sebaliknya. Hasil yang diharapkan adalah dapat menggeneralisasi nilai rating yang didapat dari sekumpulan ulasan konsumen.

II. Pendahuluan

Semakin banyaknya media sosial yang digunakan membuat konsumen memiliki banyak kesempatan untuk memberi ulasan produk kecantikan yang mereka gunakan. Terutama untuk produk obat jerawat, produk ini merupakan salah satu produk yang cukup sering digunakan oleh konsumen. Banyak dari mereka berbagi pengalaman yang mereka alami setelah menggunakan suatu produk. Contohnya pada website *female daily*, pada website tersebut terdapat banyak produk yang dapat diberi ulasan oleh konsumen dan konsumen lain dapat membaca ulasan tersebut. Pada website ini sudah terdapat rating yang diberikan oleh konsumen setiap memberikan ulasan, namun rating yang diberikan bersifat personal. Konsumen memiliki pandangan sendiri untuk menentukan nilai rating yang mereka gunakan. Hal ini menyebabkan bahwa nilai rating tidak bisa menjadi patokan yang pas untuk menentukan nilai atau kualitas dari suatu produk.

Dari masalah yang didapat, maka dibutuhkan sistem yang dapat menggeneralisasi nilai rating sesuai dengan ulasan yang diberikan. Selain untuk menggeneralisasi nilai rating, tujuan lainnya adalah untuk memberi nilai rating kepada ulasan yang tidak memiliki rating yang banyak tersebar di internet. Hal ini akan memudahkan produsen dalam menentukan kualitas dari produk yang mereka pasarkan hanya dengan melihat nilai rating dari produk tersebut.

Hasil ulasan yang diunggah konsumen dapat digunakan sebagai data untuk melakukan proses klasifikasi teks menjadi rating. Klasifikasi teks atau kategorisasi teks merupakan proses yang secara otomatis menempatkan dokumen teks ke dalam suatu kategori berdasarkan isi dari teks tersebut. Dalam kasus ini data yang digunakan untuk membangun sistem adalah data yang terdapat dalam website *female daily* yang membahas mengenai obat jerawat, topik ini dipilih karena obat jerawat merupakan salah satu produk kecantikan yang banyak digunakan oleh konsumen wanita. Ulasan yang diambil sudah memiliki kelasnya masing-masing, kelas disini direpresentasikan sebagai nilai rating yang memiliki rentang nilai antara 1-5.

Sistem yang akan dibangun untuk dapat mengklasifikasikan teks menjadi bentuk rating dibangun dengan menggunakan metode *Long Short-Term Memory* (LSTM). Metode ini dipilih

karena data yang digunakan berbentuk sekuens dan LSTM memiliki performansi yang baik jika digunakan untuk data yang berbentuk sekuens seperti data teks. Input yang akan diterima oleh sistem berupa data teks ulasan produk dan output yang akan dihasilkan berupa nilai rating dengan rentang nilai 1-5.

III. Dataset

Dataset yang digunakan untuk membangun sistem diambil dari website *female daily* yang membahas mengenai obat jerawat. Topik tersebut dipilih karena banyak konsumen yang telah memberikan ulasan pada produk tersebut. Terdapat kurang lebih 10 produk obat jerawat yang diambil ulasannya sebagai dataset. Dari setiap produk diambil kurang lebih 10 ulasan.

Dataset yang digunakan berjumlah 200 ulasan dengan nilai rating yang bermacam-macam, dataset tersebut selanjutnya dibagi menjadi tiga data yaitu sebesar 100 data sebagai data latih, 50 data sebagai data validasi, dan 50 data sebagai data uji. Data latih digunakan untuk melatih model, data validasi digunakan untuk mengecek performansi model pada saat pelatihan, sedangkan data uji digunakan untuk mengecek performansi model setelah selesai dilatih.

Contoh potongan dataset yang digunakan dapat dilihat pada tabel. Terdapat lima ulasan dengan nilai rating yang berbeda-beda, rating dan ulasan yang diberikan oleh konsumen berdasarkan pandangan mereka sendiri.

Ulasan	Rating
jerawat tetep aja ga kempes kempes. mahalnya aja. kemasannya juga ribet kalo mau buka (mungkin skrg ada kemasan baru?). sekarang ku beralih ke acnol lotion. itu baru bikin kering jerawat. bahkan bruntusan!	1
dulu sempat aku pakai ini denger denger katanya bagus buat kempesin jerawat, aku pakai ini tapi jerawat kecil besoknya muncul jerawat yang lainnya. mungkin aku salah cara pakainya kali ya, mana udah beli 2 kali juga belum ada perubahan	3
Tea Tree Oil ini persis kaya temen. Kadang baik, kadang ngeselin. Baiknya karena menenangkan jerawat gue. Ngeselinnya karena suka memicu jerawat lain yang tadinya engga ada dimuka gue dan spotnya itu deketan. Pasang-surut aku sama kamu	5

IV. Analisis Fitur dan Metode Klasifikasi serta Eksperimen

1. Analisis Fitur

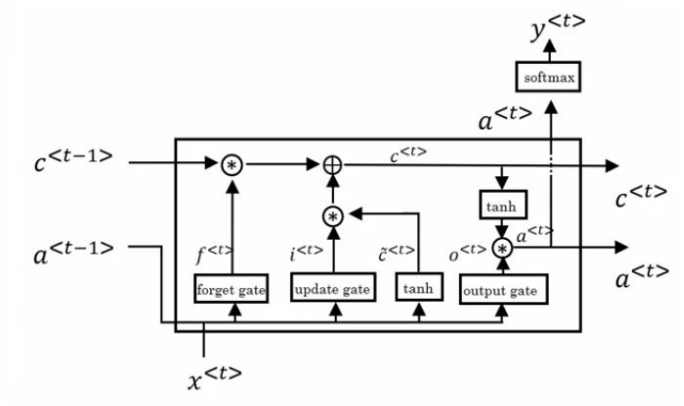
Pada sistem ini tahap ekstraksi fitur dilakukan menggunakan metode *embedding*. *Embedding* merupakan proses inisialisasi vektor kata yang bertujuan untuk menentukan kemiripan antar kata yang direpresentasikan dalam bentuk vektor. Dengan melakukan tahap ini maka fitur yang diambil berbentuk vektor kata yang berupa numerik. Karena *embedding* dapat menentukan kemiripan antar kata, maka kamus kata yang digunakan

akan lebih sedikit karena kata-kata yang memiliki tingkat kemiripan yang tinggi akan dianggap menjadi satu makna kata. Contoh dari penggunaan embedding dapat dilihat pada tabel

input	[3,5,16,0]
output	[[0.32,0.45,...,0.12],...,[0.21,0,...,0.7]]

2. Algoritma Klasifikasi

Untuk mengimplementasikan sistem yang akan dibangun digunakan pendekatan *deep learning*, metode yang dipilih adalah metode *Long short-Term Memory* (LSTM). LSTM merupakan bagian dari *Recurrent Neural Network* (RNN), perbedaan LSTM dan RNN terletak pada modul yang dimiliki dan masukan yang digunakan. Di dalam modul LSTM terdapat empat komponen, yaitu *memory cell*, *update gate*, *forget gate*, dan *output gate*. Kunci utama dari LSTM adalah *memory cell*, dengan adanya *memory cell* maka model akan dapat menyimpan ingatan lebih banyak dibandingkan dengan model RNN biasa.



$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$
$$\Gamma_u = \sigma(W_u[\, a^{<t-1>}, x^{<t>}] + b_u)$$
$$\Gamma_f = \sigma(W_f[\, a^{<t-1>}, x^{<t>}] + b_f)$$
$$\Gamma_o = \sigma(W_o[\, a^{<t-1>}, x^{<t>}] + b_o)$$
$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$
$$a^{<t>} = \Gamma_o * \tanh c^{<t>}$$

Gambar di atas merupakan ilustrasi dari satu modul LSTM. Terdapat fungsi tanh dan sigmoid, fungsi sigmoid digunakan pada ketiga *gate*. Masukan yang diterima oleh satu modul LSTM adalah $c^{<t-1>}$ dan $a^{<t-1>}$, dimana nilai c merepresentasikan nilai dari *memory cell* yang didapat pada modul sebelumnya dan nilai a merepresentasikan nilai keluaran yang sebenarnya dari modul sebelumnya. Rumus yang tertera adalah rumus yang digunakan untuk menghitung nilai masukan yang terdapat pada setiap *gate*.

3. Library Pendukung

Library yang digunakan untuk membangun sistem menggunakan metode LSTM adalah sebagai berikut:

- Keras, digunakan untuk membangun model dengan metode LSTM
- Tensorflow, merupakan *backend* dari keras
- Matplotlib, digunakan untuk memvisualisasi hasil dari klasifikasi
- Pandas, digunakan untuk membaca file dalam excel

4. Tahap Pra proses

Tahap praproses yang dilakukan sebelum dataset dimasukkan kedalam model LSTM adalah sebagai berikut:

- ***Case Folding*:** pengkonversian teks menjadi bentuk standar, yaitu merubah semua huruf kapital menjadi huruf kecil
- ***Punctuation Removal*:** membersihkan dataset dengan menghilangkan semua tanda baca
- ***Tokenizing*:** memecah suatu kalimat menjadi potongan-potongan kata yang berdiri sendiri, kemudian potongan kata tersebut direpresentasikan menjadi bentuk numerik
- ***Pad Sequence*:** menyamakan panjang kata sesuai dengan yang ditentukan dan menambahkan nilai 0 untuk kalimat yang lebih pendek dari panjang yang ditentukan

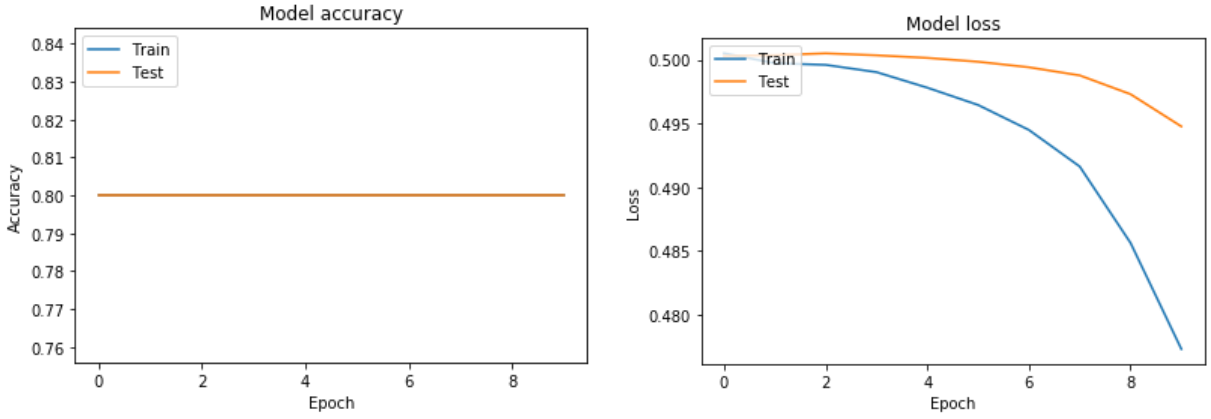
5. Parameter

Parameter yang digunakan untuk membangun sistem adalah sebagai berikut:

- **Ukuran matrix embedding yang digunakan sepanjang 32:** embedding_size = 32
- **Model LSTM yang dibangun memiliki 2 layer, bersifat bidirectional:** model.add(Bidirectional(LSTM(units=32, dropout=0.2, return_sequences=True, recurrent_dropout=0.2, activation='tanh')))
- **Fungsi aktivasi menggunakan fungsi softmax:** model.add(Dense(5, activation=softmax))
- **Loss function menggunakan model binary:** model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

V. Evaluasi dan Analisis Hasil Klasifikasi

Hasil akurasi model memiliki akurasi yang cukup baik, dapat dilihat pada grafik model accuracy bahwa data validasi dan data latih selalu berimpit dan konstan pada titik 0.80.



Pada grafik model loss, dapat dilihat bahwa nilai loss menurun seiring dengan bertambahnya epoch. Tetapi jika epoch terlalu besar, loss pada data latih akan menurun

dan loss pada data tes bertambah. Hal tersebut akan membuat model menjadi overfit dan tidak dapat menggeneralisir dengan baik.

VI. Kesimpulan

Kesimpulan yang didapat adalah hasil dari performansi adalah akurasi sebesar 80% pada data test. LSTM merupakan salah satu metode yang memiliki performansi yang cukup baik bila digunakan untuk data yang berbentuk sekuens. Kesulitan dalam membuat sistem klasifikasi ulasan produk menjadi bentuk rating adalah tahap pembersihan data. Kebanyakan data ulasan produk bersifat non-formal. Sehingga terdiri dari banyak kata singkatan, pencampuran kata indonesia dan inggris, perbedaan penulisan kata, dan lainnya. Sehingga menyebabkan model membutuhkan data yang banyak agar dapat mempelajari data tersebut.

VII. Referensi

Zhang, X.F, Huang, H.Y, Zhang K.L. 2009. KNN Text Categorization Algorithm Based on Semantic Centre. 2009 International Conference on Information Technology and Computer Science

Sepp Hochreiter and J ¨urgen Schmidhuber. Long short-term memory.Neural computation, 9(8):1735–1780, 1997

https://indoml.com/2018/04/13/pengenalan-long-short-term-memory-lstm-dan-gated-recurrent-unit-gru-rnn-bagian-2/

- **Pembagian Tugas**

Tugas/Nama	Aditya Alif Nugraha	Nadine Azhalia
Membangun Dataset	50%	50%
Membuat Laporan	30%	70%
Membuat Program	70%	30%