Laporan Tugas Besar
Klasifikasi Sentimen Pada Komentar di Halaman Web Zomato

Oleh:

I Putu Indra Aristya	1301154219
Yogi Wisesa Chandra	1301154282
Raginda Firdaus	1301150037

Universitas Telkom

Bandung

2018

Abstrak

Klasifikasi sentimen merupakan hal yang dapat membedakan sentimen pada suatu kalimat. Perbedaan sentimen yang umum biasanya adalah sentimen positif dan negatif. Namun, terkadang penilaian sentimen dan *rating* yang diberikan oleh pengguna tidak sesuai. Pada penelitian ini dicoba untuk melakukan klasifikasi sentimen positif dan negatif pada data komentar pada restoran – restoran yang ada di halaman web Zomato.com dengan menggunakan metode klasifikasi *Multinomial Naive Bayes* dan *Support Vector Machine* dengan fitur *Mutual Information* pada 75 data dengan data latih 125 data. Hasil yang didapatkan adalah nilai akurasi sebesar 82.67% dengan menggunakan *Multinomial Naive Bayes* pada data yang sudah dilakukan *stemming* menggunakan *library* Sastrawi.

Kata kunci: sentimen, multinomial naive bayes, support vector machine, mutual information, zomato

Bab I : Pendahuluan

A. Latar Belakang

Pada masa ini, aplikasi atau halaman web penyedia informasi sudah semakin berkembang dan banyak digunakan. Contoh informasi yang disediakan adalah informasi tempat wisata, hotel, harga tiket pesawat atau kereta dan lainnya. Penyedia informasi tersebut juga menambahkan kolom komentar atau *review* yang dapat diisi oleh setiap orang yang seharusnya sudah pernah memiliki pengalaman menggunakan barang atau jasa dari informasi yang disediakan. Komentar atau *review* tersebut juga dilengkapi dengan *rating* yang dapat mengelompokkan komentar – komentar dan penilaian dengan angka dari suatu produk atau jasa. Namun, terkadang ada saja komentar yang diberikan tidak sesuai dengan *rating* yang diberikan. Contohnya, komentar terhadap suatu hotel sudah sangat bagus dari segi pelayanan, kebersihan dan ruangan namun nilai yang diberikan hanya 3 dari 5, yang seharusnya bisa mencapai nilai 4 ataupun 5.

Oleh karena itu, sistem klasifikasi sentimen dari suatu komentar dapat berguna pada masalah ini. Dari komentar yang diberikan akan langsung ditentukan apakah sentimen atau nilai dari komentar tersebut. Pada kasus ini, kami mengambil contoh komentar pada halaman web atau aplikasi Zomato (penyedia informasi restoran – restoran). Dari komentar tersebut akan diklasifikasikan kelas sentimen positif (1) atau negatif (0). Sistem klasifikasi dibuat dengan bantuan penerapan ilmu *natural language processing* dan akan menghasilkan sebuah model yang paling cocok dengan kasus sentimen komentar.

B. Tujuan

Penelitian ini memiliki tujuan untuk dapat membangun model *machine learning* yang cocok digunakan untuk klasifikasi sentimen pada komentar yang ada di halaman web Zomato berbahasa Indonesia.

Bab II : Data

Pada penelitian ini, data diambil dengan melakukan *scrapping* pada halaman web Zomato dengan mengambil komentar – komentar pada setiap *review* restoran. Data yang diambil berjumlah 200 data dengan masing – masing kelas adalah 100 data. Dengan itu, digunakan 150 data sebagai data latih dan 50 data sebagai data validasi.

Data yang diambil berupa kalimat – kalimat dengan bahasa Indonesia yang semi-formal. Terdapat komentar yang menggunakan bahasa Indonesia yang formal (sesuai EYD) dan non-formal (menggunakan istilah sehari – hari). Pada kalimat yang menggunakan bahasa non-formal ataupun terdapat singkatan dapat mempersulit proses klasifikasi karena bisa saja kata yang sebenarnya memiliki arti sama dianggap berbeda, contohnya pada kalimat pertama terdapat kata "juga" dan kalimat kedua terdapat kata "jg". Selain itu, pemilihan kata yang digunakan banyak terdapat pengulangan huruf dan angka, seperti "cekernyaaa pedes tapi enakkkkk. tempatnya kurang gede tapi nyaman. harganya ga terlalu mahal dan porsinya banyakkk. menu nya juga ga cuma ceker tapi ada yang lain juga seperti wings, dll" dan "Tempatnya unik dan kebetulan waktu pas dtg kesini lagi agak sepi, jadi enak2 aja lama2 disini. Porsi makanannya ngenyangin juga. Buat yang suka pasta, bisa mampir kesini.".

Data dengan kelas negatif dan positif juga relatif berbeda. Pada data yang didapatkan, rata – rata jumlah kata untuk kelas negatif adalah 81,59 dan untuk kelas positif adalah 66,25. Hal tersebut karena, biasanya konsumen yang merasakan pengalaman yang buruk atau memberikan *review* yang buruk menggunakan kata yang lebih banyak daripada *review* positif untuk meluapkan pengalamannya.

Bab III : Metode Penelitian

Penelitian ini menggunakan metode *Count Vectorizer* dan *Mutual Information* sebagai ekstraksi fitur dan *Multinomial Naivè Bayes* (NB) dan *Support Vector Machine* (SVM) sebagai metode klasifikasi

Pre-processing yang digunakan pada data ini adalah menghapus tanda baca dan mengubah semua huruf menjadi huruf kecil. Tujuannya adalah pada kasus sentimen kita tidak memerlukan informasi yang didapat dari tanda baca dan pada saat melakukan perubahan kata menjadi *vector* kata tidak dianggap berbeda dengan berbedanya 1 huruf saja. Misalnya, pada saat melakukan *count vectorizer* kata "Saya" dan "saya" dianggap berbeda karena sensitif terhadap huruf, maka perlu diubah menjadi huruf kecil agar kata dianggap sama. *Pre-processing* juga dilakukan dengan melakukan *stemming* menggunakan *library* Sastrawi. *Stemming* akan mengubah kata – kata yang ada, menjadi kata – kata dasar sesuai dengan korpus kata yang ada pada *library* Sastrawi tersebut. Data yang sudah melewati proses *pre-processing* sudah dapat digunakan untuk klasifikasi. Pada proses klasifikasi menggunakan *multinomial naivè bayes* (MNB), data yang diberikan masih berupa kalimat – kalimat yang dihitung probabilitas kata terhadap suatu kelas(1) dan probabilitas kelas tersebut. Dengan nilai probabilitas tersebut kemudian dihitung probabilitas suatu kalimat terhadap kelas(2) dan dipilih hasil probabilitas yang lebih tinggi.

$$P(kata|kelas) = \frac{count(kata\ dalam\ kelas) + 1}{count(semua\ kata\ dalam\ kelas) + count(kata\ unik)} \quad (1)$$

$$P(kalimat|kelas) = \sum \log\big(P(kata|kelas)\big) \quad (2)$$

Klasifikasi yang kedua dilakukan menggunakan *support vector machine* (SVM). Perbedaannya, pada penggunaan SVM ini data yang dimasukkan adalah data yang telah dipreproses dengan *count vectorizer* dan diambil fiturnya dengan *mutual information*. *Mutual Information* (MI) akan menentukan kemiripan antara suatu kata dengan suatu kelas dan dilakukan pengurutan nilai MI dari yang terbesar ke terkecil. Kemudian, akan dipilih jumlah nilai terbesar tersebut untuk dipilih menjadi fitur untuk diklasifikasi menggunakan SVM.

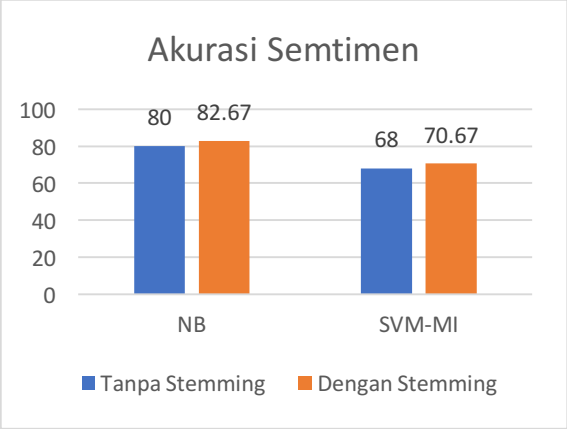
Bab IV : Skenario dan Analisa

A. Skenario

Pada penelitian ini akan digunakan skenario penggunaan *stemming* dan tanpa dilakukan *stemming*. Parameter – parameter lainnya yang di*tuning* adalah parameter jumlah fitur MI yang diambil. Namun, penentuan jumlah fitur MI yang diambil sudah ditentukan sebelumnya dengan mencoba – coba hingga mendapatkan akurasi yang cukup bagus. Maka, dibandingkan nilai akurasi yang dihasilkan pada data yang dilakukan *stemming* dan tidak dilakukan *stemming.*

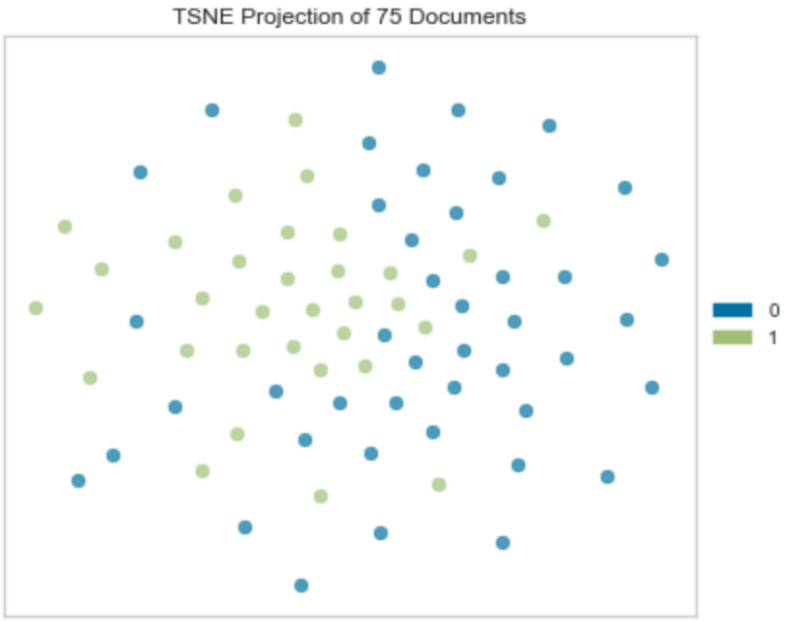
B. Analisa

Pada proses pengujian hasil yang didapatkan adalah sebesar 80% dengan menggunakan metode *Multinomial Naive Bayes* tanpa *stemming* dan 82,67% dengan menggunakan *stemming*. Kemudian, saat menggunakan *Support Vector Machine* kernel linear, didapatkan akurasi sebesar 68% tanpa *stemming* dan 70,67% dengan menggunakan *stemming*.



Dengan itu, didapatkan kesimpulan bahwa dengan melakukan *stemming* Sastrawi, akurasi yang didapatkan dengan menggunakan MNB dan SVM lebih besar daripada tidak dilakukan *stemming*. Pada hal ini, *stemming* akan membantu untuk mengubah kata – kata menjadi kata dasar sehingga memiliki arti yang sama dan lebih mudah untuk dilakukan klasifikasi dan pengambilan fitur atau bisa diartikan dapat mengurangi variansi kata yang ada sehingga memudahkan klasifikasi.

Dengan menggunakan SVM didapatkan hasil yang kurang bagus karena kemungkinan persebaran data yang kurang baik. Pada penelitian ini hanya digunakan kernel *linear* yang menurut visualisasi data uji menggunakan TSNE didapatkan visualisasi seperti Gambar 1.



Gambar 1 Visualisasi Data Uji

Selain itu, pada hasil klasifikasi yang tidak sesuai dengan label disebabkan karena ada data yang berbahasa Inggris sedangkan data dominan adalah data bahasa Indonesia. Maka dengan MNB sulit diklasifikasi karena terdapat perbedaan yang mencolok pada bahasa. Selain itu, ada kata yang berulang yang bisa menyebabkan kesalahan prediksi, seperti pada kalimat "gue kesini sama my bf sepi banget parahsihh itu hari jumat sore gitu gue pilih di rooftop nya gue cuma pesen cumi goreng tepung sm es jeruk cumi nya benerbener keras dan es jeruknya tuh sirup coiii kebayang kan manisnya fake begitu" dan pada pengolahan teks sangat sensitif terhadap huruf (berbeda satu huruf, dianggap kata yang berbeda). Maka, pada proses *pre-processing* dapat diubah ke bahasa formal terlebih dahulu.

C. Kesimpulan

Pada penilitian ini, dapat ditarik kesimpulan bahwa proses *pre-processing* merupakan proses yang cukup penting untuk dilakukan. *Pre-process* yang dianjurkan

adalah proses *stemming* dan juga mengubah kata – kata yang tidak formal menjadi formal. Proses *stemming* terbukti memberikan hasil yang lebih baik daripada tidak dilakukan *stemming* dengan mengurangi variansi kata yang ada sehingga memudahkan klasifikasi. Fitur *Mutual Information* dapat membantu menaikkan akurasi dengan mengurangi fitur yang tidak penting pada proses klasifikasi. Kesulitan yang didapatkan saat membuat penelitian ini adalah pengambilan data yang dilakukan dengan *scrapping* pada halaman Web Zomato yang labelnya tidak dapat diambil. Maka diperlukan waktu untuk melakukan *labeling* data.

Kontribusi.

No	Nama	NIM	Bagian Kerja	Persentase
1	I Putu Indra Aristya	1301154219	*Scrapping data awal*, *code Multinomial Naive Bayes, code MI, dan laporan*	(100/3)%
2	Yogi Wisesa Chandra	1301154282	*Code SVM, CountVectorizer, stemming, scrapping tambahan data, code convert emoji*	(100/3)%
3	Raginda Firdaus	1301150037	*Mencoba code kNN (tidak dimasukkan karena hasil tidak lebih bagus dari SVM dan MNB), labeling data, code cleaning data*	(100/3)%