PROYEK *TEXT ANALYTICS*

Analitika Media Sosial

Laporan Akhir

Adrika Novrialdi - 1606877585

Alif Ahsanil Satria - 1606882540

Naomi Riana - 1606889540

Deskripsi Tugas

Proyek *text analytics* yang dikerjakan adalah merancang model untuk mengklasifikasi *tweet* berbahasa Indonesia. *Tweet* akan digolongkan berdasarkan sentimennya (positif, negatif, atau netral). Model klasifikasi akan diuji dengan *test set* yang berisi 8000 *tweets*. Klasifikasi dijalankan dengan bantuan *train set* yang sudah disediakan sebanyak 3462 *tweets.*

Metodologi

Secara umum terdapat terdapat tiga himpunan data yang digunakan, *training data, tester data,* dan *testing data*. *Training* data berisi *tweets* berlabel yang akan dijadikan acuan dalam klasifikasi *tester data. Training data* yang akan menentukan kualitas klasifikasi. *Tester data* merupakan salah satu himpunan data yang akan diuji dalam klasifikasi, dalam ini untuk menguji jalan awalnya model. Sementara untuk data yang sebenarnya akan diklasifikasi adalah *testing data.* Perbedaan dua himpunan data hanyalah *tester data* memiliki jumlah yang lebih sedikit sehingga cocok untuk digunakan untuk pengecekan awal. *Testing data* kemudian dipecah melalui tokenisasi berdasarkan karakter spasi menjadi fitur. Fitur - fitur dari *tester data* yang ditokenisasi melalui proses normalisasi dengan membuang *punctuation,* digit, *extra space,* dan huruf berulang. Fitur dari *tweet* juga akan di-*stem* menjadi kata dasar dari fitur tersebut. Fitur - fitur tersebut diseleksi kembali untuk membuang fitur yang merupakan *stopwords* berdasarkan kamus yang telah dibuat.

Fitur - fitur yang sudah melalui proses - proses tadi diekstraksi dalam bentuk *Bag-of-Words*. Klasifikasi dilakukan dengan *Multinomial Bayes*, yaitu menghitung besar kecenderungan sentimen *tweet* berdasarkan fitur/kata yang muncul pada *train set*. Sentimen fitur/kata dapat dilihat berdasarkan kamus kata - kata positif dan

negatif yang sudah disediakan. Hasil dari klasifikasi dengan *Multinomial Bayes* akan menjadi dasar untuk pemberian label sentimen (0 untuk negatif, 1 untuk positif) untuk *testing data*. Kami juga melakukan pendekatan fitur leksikon sentimen yang dilakukan pada corpus testing dalam memprediksi sentimen.

Metode/Pendekatan

Secara umum, kami mengekstraksi fitur dengan dua pendekatan, yaitu *bag of words* dan leksikon sentimen. Pada fitur leksikon sentimen, yang kami lakukan adalah menghitung banyaknya kata yang mengandung sentimen positif dan negatif pada masing - masing *tweet* di corpus testing. Apabila di sebuah *tweet* lebih banyak kata bersentimen positif, maka prediksi sentimen nya adalah positif (begitu juga sebaliknya).

Untuk bagian bag of words, kami melakukan tiga pendekatan untuk mengekstraksi bag of words nya. Pendekatan pertama adalah dengan stemming. Pada pendekatan ini, langkah yang dilakukan adalah normalisasi, menghapus stopword, dan stemming pada corpus training dan testing, kemudian diekstraksi bag of words terhadap 5000 vocabulary yang paling sering muncul pada kedua corpus tersebut. Pendekatan kedua adalah tanpa stemming. Pada pendekatan ini, langkah yang dilakukan adalah normalisasi, menghapus stopword, dan normalisasi lanjutan. Untuk melakukan normalisasi lanjutan, yang kami lakukan adalah mengekstraksi vocabulary beserta count nya dari hasil normalisasi + stopword removal, lalu melakukan normalisasi 50-60 kata yang paling sering muncul pada corpus training. Kami juga melakukan ekstraksi vocabulary lanjutan khusus *adjective* saja, lalu kembali melakukan normalisasi 50-60 kata *adjective* yang paling sering muncul pada corpus training. Hasil mapping normalisasi dari kedua

hal tersebut disimpan pada file normalisasi_mapping.txt. Hasil mapping normalisasi ini diterapkan pada corpus training dan testing, lalu diekstraksi bag of words terhadap 7000 vocabulary yang paling sering muncul pada kedua corpus tersebut. Pendekatan ketiga adalah pendekatan kedua yang dilanjutkan dengan normalisasi dengan sinonim yang daftar kata bersinonim nya terdapat pada file thesaurus.json. Pada pendekatan ini, kata *adjective* pada thesaurus.json yang saling bersinonim akan di-*mapping* ke makna yang sama. Kemudian, hasil mapping ini diterapkan pada corpus training dan testing, lalu diekstraksi bag of words nya pada 4000 vocabulary yang paling sering muncul pada kedua corpus tersebut. Sebagai catatan, setelah bag of words diekstraksi, bagian training dan testingnya dipisah. Selain itu, karena proses stemming pada corpus training dan testing memakan waktu yang cukup lama, maka untuk pendekatan yang pertama kami sudah menyimpan tweet yang sudah cleaned pada file cleaned_stemming.csv untuk training dan cleaned_stemming_test.csv untuk testing.

Eksperimen dan Hasil

Eksperimen dilakukan dengan menggunakan *tester data* sebanyak sepuluh *tweet*. Eksperimen berikutnya dilakukan dengan *testing data* sebanyak 8000 *tweet*. Pada pendekatan leksikon sentimen, kami mendapatkan akurasi sekitar 65-67% (tidak di-*screenshot*). Lalu, pada pendekatan pertama dengan fitur bag of words, kami mendapatkan akurasi sebesar 85,79%

.:Informasi:.

Uploaded File: test_result.csv
Type: application/vnd.ms-excel
Size: 61.416015625 KB
Total Benar 1600
Accuracy: 85.790884718499

*jika terjadi error terkait 'mysql', coba unggah sekali lagi.

See Current RankingsError submiting record: INSERT command denied to user 'zprot2whsf79zkfv'@'ec2-35-171-85-105.compute-1.amazonaws.com' for table 'submission_logs'

Kemudian, pada pendekatan kedua dengan bag of words, kami mendapatkan akurasi sebesar 87,02% (akurasi terbesar)

4	Regenmagier	70.00	87.02

Terakhir, pada pendekatan ketiga dengan bag of words, kami mendapatkan akurasi sekitar 84,60% (tidak di-*screenshot*)

Analisis *Error*

Untuk pendekatan fitur leksikon sentimen, penyebab pertama terjadinya misklasifikasi adalah tidak semua kosa kata pada *corpus testing* terdapat pada leksikon sentimen yang kami miliki. Penyebab lainnya adalah mekanisme yang kami lakukan pada pendekatan ini tidak memperhatikan konteks kalimat secara semantik dalam memprediksi sentimen, hanya memperhatikan sentimen kata demi kata pada sebuah *tweet*.

Sementara penyebab misklasifikasi dengan pendekatan *bag of words* (ketiga pendekatan), yang pertama adalah tidak semua kosa kata yang bisa berpengaruh terhadap sentimen sebuah *tweet* bisa ternormalisasi sehingga dalam penghitungan probabilitas berdasarkan *multinomial naive bayes*. *Tweet* yang seharusnya kental dengan sentimen positif/negatif probabilitasnya bisa sangat rendah karena probabilitas satu atau beberapa kata yang sangat kecil. Penyebab lainnya adalah tidak semua *stopword* pada *corpus training* dan *testing* bisa dihilangkan (*stopword* dalam hal ini adalah kata yang sebenarnya tidak mempengaruhi sentimen, tapi bisa saja membuat sentimennya menjadi condong ke salah satu (positif atau negatif) karena tidak meratanya distribusi kata tersebut pada corpus training yang bersentimen positif/negatif). Sebagai contoh, kata “saya” memiliki sentimen netral, tetapi bisa saja membuat prediksi sentimen menjadi positif apabila distribusi kata

“saya” pada *corpus training* bersentimen positif jauh lebih banyak daripada yang negatif (begitu juga sebaliknya). Penyebab lainnya adalah mekanisme yang kami lakukan pada pendekatan ini tidak memperhatikan konteks kalimat secara semantik dalam memprediksi sentimen, hanya memperhatikan kata demi kata untuk memprediksi sentimen (karena *language model* yang kami pakai adalah *unigram*). Untuk yang spesifik pada pendekatan ketiga di bag of words, penyebabnya adalah ambiguitas kata. Sebagai contoh, misalkan terdapat kata “bisa” pada salah satu tweet di corpus testing yang berkalimat “dia terkena bisa ular”. Ternyata, kata “bisa” pada corpus training di-*mapping* ke makna yang bersinonim dengan “mampu”, “dapat”, dan sejenisnya. Alhasil, kalimat “dia terkena bisa ular” kemungkinan besar sentimennya adalah positif, padahal seharusnya negatif.

Kesimpulan dan Saran

Berdasarkan hasil di atas, pendekatan kedua pada bag of words memberikan akurasi yang paling besar di antara pendekatan lainnya. Model klasifikasi *tweet* berbahasa Indonesia ini diharapkan dapat menganalisis *tweet* dari konteks semantik. Dikarenakan keterbatasan kemampuan, akhirnya kami hanya menganalisis sentimen dari suatu tweet berdasarkan kata per kata. Perbaikan lain yang bisa dilakukan kedepannya adalah proses normalisasi kata yang lebih kaya lagi sehingga cakupan kata yang bisa dinormalisasi bisa sebanyak mungkin.