IMPLEMENTASI NAÏVE BAYES CLASSIFIER UNTUK KLASIFIKASI PENCARIAN TEMPAT KOST

Taryadi, Slamet Joko Prasetiono
STMIK Widya Pratama
tari_ball@stmik-wp.ac.id, blackjack@gmail.com

RINGKASAN

Artikel ini membahas penggunaan Naïve Bayes Classifier dengan menggunakan data dari web www.jakarta.craiglist.org. Craiglist menyediakan iklan baris dan forum lokal untuk pekerjaan , perumahan untuk dijual, personal, layanan, komunitas lokal dan acara. Dengan menggunakan aplikasi *web-crawling* dengan berbasis naïve bayes, akan digunakan untuk mengumpulkan data teks dari bagian penyewaan tempat kost di wilayah Jakarta dan mengklasifikasikan setiap posting ke kota mana termasuk teks pada judul iklan di Craigslist. Dengan menggunakan data yang sama, berupaya untuk menggunakan judul iklan untuk menentukan braket harga yang termasuk di dalamnya. Sebagai pengolah *naïve bayes classifier* digunakan versi open source yang tersedia di https://github.com/alexandru/stuff-classifier untuk melakukan klasifikasi. Membandingkan naïve bayes dengan statistik Tf-Idf. Berdasarkan hasil pengolahan menunjukkan bahwa judul saja bukan indikator yang baik tentang berapa kisaran harga suatu iklan. Selain itu, naïve bayes melakukan sedikit lebih baik dibandingkan metode Tf-Idf yang hanya berdasarkan pada kebenaran klasifikasi. Untuk mendapatkan hasil yang lebih baik, dilakukan penyujian *web crawler* lebih lanjut dan memasukkan deskripsi tekstual yang lebih panjang dari setiap daftar ke dalam proses klasifikasi. Dengan melakukan hal tersebut ternyata tidak menghasilkan perubahan perilaku output dari salah satu classifier.

Kata Kunci: naïve bayes, klasifikasi, web crawler

1. PENDAHULUAN

Mahasiswa sebuah perguruan tinggi biasanya mengalami masalah dalam menemukan tempat kost yang sesuai, murah dan terjangkau. Salah satu sumber yang digunakan oleh mahasiswa untuk mencari tempat kost adalah jakarta.craigslist.org. web Craigslift menyediakan pencarian lokasi, menyewakan, dan mencari teman sekamar sebagai salah satu fitur dari iklan yang diposting dalam penyediaan kebutuhan hidup.

Salah satu hal yang paling sulit, ketika dihadapkan dengan daftar tempat kost adalah untuk menentukan apakah deskripsi yang tersedia pada daftar iklan memiliki nilai yang sesuai dengan kebutuhan dari mahasiswa.

Selain harga, sering kali fitur pencarian Craigslist bukan jumlah hasil pencarian pada area tertentu. Misalnya, mencoba mencari tempat kost di daerah Jakarta, maka pencari akan melakukan pencarian dengan kata kunci Jakarta sebagai kata utama, tetapi hasil yang didapatkan ternyata Jakarta dan kota yang mendekati kata Jakarta seperti Yogyakarta juga dimungkinkan untuk ditampilkan.

Secara sederhana, classifier Naïve Bayes menggunakan keberadaan fitur dalam data untuk menentukan kelas dari setiap instance

tertentu (Artaye, 2015). Diberikan data pelatihan, pengklasifikasi mengkorelasikan keberadaan fitur terhadap kelas tempat instance pelatihan itu berada. Setelah melatih sejumlah data, classifier dapat menggeneralisasi setiap instance data tertentu yang dilihatnya menggunakan frekuensi fitur yang telah diamati dalam set pelatihan (Prasetyo, 2014).

Dalam beberapa hal, klasifikasi yang tepat tergantung pada domain data itu sendiri (Felida, 2017). Dengan sejumlah fitur yang terbatas dan sampel data apa pun, ada kemungkinan bahwa classifier akan dapat melakukan klasifikasi dengan benar setelah melihat informasi pelatihan yang cukup (Nurjoko, 2016). Dengan beberapa pembelajaran yang diawasi, instance baru dari setiap kelas dapat digunakan untuk melatih *classifier* lebih lanjut sehingga menghasilkan perkiraan fitur yang lebih baik yang menentukan kelas (Kusrini, 2009). Namun, untuk kumpulan data seperti bahasa Inggris seringkali sulit untuk menilai; karena jumlah fitur sangat besar (seluruh bahasa Indonesia).

Upaya dapat dilakukan untuk mengurangi kompleksitas dataset ini melalui lemmatization, namun dalam contoh, saya telah melihat, Naïve Bayes memiliki kinerja yang baik dalam menangani subset bahasa yang lebih kecil, seperti filter dan detektor spam. Studi kasus ini bermaksud untuk mencoba menggunakan subset tertentu dari bahasa Inggris yang dikaitkan dengan jargon berburu tempat kost dan menemukan indikasi nilai dalam kata-kata ini.

2. METODE PENELITIAN

Penelitian ini mengunakan metode ekperimenta dengan melakukan pengukuran data yang dikumpulkan melalui situs jakarta.craigslist.org. Tahapan yang dilakukan dalam penelitian ini adalah sebagai berikut:

2.1. Pengumpulan Data

Sumber data penelitian ini menggunakan berasal dari situs jakarta.craigslist.org yang merupakan situs yang menyediakan pencarian lokasi, menyewakan, dan mencari teman sekamar sebagai salah satu fitur dari iklan yang diposting dalam penyediaan kebutuhan hidup.

2.2.Pemodelan dan Analisis

Untuk melakukan pemodelan data digunakan dengan melakukan perayapan web dengan menggunakan perangkat open source yang tersedia di https://github.com/alexandru/stuff-classifier. Model klasifikasi menggunakan model Naïve Bayes Classifier dan Tf-Idf (Term Frequency - Inverse Document Frequency) (Feldman, 2007)(Melita, dkk., 2018) dan saling dibandingkan hasil perayapan data di situs yang dituju.

3. HASIL DAN PEMBAHASAN

3.1.STRUKTUR CRAWLER DAN WEB

Informasi tentang penyewaan tempat kost dapat ditemukan di http://jakarta.craigslist.org setelah memeriksa halaman ini, dapat terlihat struktur data itu sendiri. Semua posting adalah lokasi di dalam blokquote tag html dengan atribut id dari baris toc. Bersarang di dalam elemen ini adalah dua jenis data. Header tag menentukan tanggal posting yang dapat diabaikan, dan iklan itu sendiri. Setiap iklan disimpan dalam tag paragraf dengan kelas baris, di dalam tag ini, lokasi tempat kost berada dalam tag kecil dan tanda kurung. Judul setiap posting terletak di antara jangkar tag pembuka dan penutup. Karena tidak ada penandaan tag yang tidak perlu di dalam satu sama lain, html dapat diuraikan dengan ekspresi reguler dan kemudian dituliskan ke file data selama perayapan. Pada bagian kedua dari studi kasus ini, kami mem-parsing

harga dari html dengan mencari tag rentang dengan kelas harga item.

Tabel 1. Hasil awal, kurang jelas

Classifier	Benar	Salah	Jumlah Data
Naïve Bayes	104	1352	485
Tf-Idf	0	1456	485

Untuk mengumpulkan data yang cukup, halaman indeks yang terletak di http://jakarta.craigslist.org/apa/ tidak cukup. Namun, pada setiap halaman adalah hyperlink ke 100 posting berikutnya. Memeriksa hasil tautan ini dalam menentukan format penggunaan craigslist URL untuk menyimpan informasi yang lebih lama. Setiap 100 posting setelah halaman indeks awal disimpan pada halaman dengan format berikut: http://jakarta.craigslist.org/apa/index (nomor halaman)00.html di mana nomor halaman adalah angka yang sama dengan atau lebih tinggi dari 1

Dengan mencatat bentuk terstruktur ini tidak hanya data tetapi URL itu sendiri, kami membuat program perayapan yang mengambil sejumlah halaman sewenang-wenang, dan menggunakan ekspresi reguler, mengumpulkan teks yang relevan ke dalam file data.

3.2.KLASIFIKASI BERDASARKAN LOKASI

3.2.1. Hasil awal tanpa pembersihan data

Setelah mengumpulkan 1456 sampel, total 811 kelas berbeda ditemukan di antara lokasi. Seperti dapat dilihat pada Tabel 1 upaya untuk mengklasifikasikan data dengan metode mana pun menghasilkan kurang dari 10% dari data yang diklasifikasikan dengan benar setelah pelatihan pada sepertiga dari dataset. Ini tidak diragukan lagi karena banyaknya lokasi yang dihasilkan oleh perayapan dan penguraian otomatis.

Menggunakan setengah dari data pelatihan alih-alih yang ketiga menghasilkan tidak ada perubahan untuk Tf-Idf atau Naif Bayes, ini menyiratkan bahwa beberapa yang benar yang diklasifikasikan Naif Bayes kemungkinan besar benar karena kebetulan dan tidak lebih. Generasi otomatis label kelas melalui teks dari hanya hasil craigslist label kelas terlalu khusus yang dibuat. Misalnya, jika sebuah iklan mencantumkan lokasi 132 Jl. Ahmad Yani, Jakarta Selatan ini akan menjadi kelas yang sama sekali baru, bahkan jika kelas Jakarta Utara ada juga. Jelas, sebagai manusia, kita tahu bahwa kedua postingan dapat diklasifikasikan sebagai berlokasi di Jakarta, dan pada kenyataannya, akan bermanfaat untuk membersihkan data kita sedemikian rupa sehingga jumlah kelas yang berbeda dipersempit dari 811 menjadi jauh. jumlah yang lebih kecil.

Tabel 2. Hasil klasifikasi setelah pembersihan data

Classifier	Benar	Tidak Benar	Jumlah Data
Naïve Bayes	203	1226	485
Naïve Bayes	231	1225	728
Tf-Idf	0	1456	485
Tf-Idf	0	1456	728

3.2.2. Pembersihan Data

Untuk membersihkan data, harus dilakukan pemetaan alamat ke kota-kota. Ini adalah tugas yang membosankan bagi manusia untuk dilakukan, tetapi tidak satu untuk Google Geocode API. Menggunakan geocode API akan memiliki akses ke maps.google.com. Permintaan sederhana untuk dilakukan, dan dengan menguraikan output untuk lokalitas alamat yang diselesaikan Google, maka dapat mempersempit daftar kelas dari 811 menjadi 503.

Meskipun jumlah ini masih banyak, perbaikan segera dalam klasifikasi data terjadi. Seperti terlihat pada Tabel 2, Naïve Bayes melihat penggandaan dari output yang diklasifikasikan dengan benar tetapi Tf-Idf masih gagal untuk mengklasifikasikan. Jumlah kelas masih terlalu tinggi untuk dapat mengurutkan setiap posting ke lokasi yang benar.

Untuk menurunkan jumlah kelas menjadi lebih mudah dikelola, dillakukan perubahan cakupan permintaan API. Sepanjang putaran pertama pembersihan digunakan specifier lokalitas untuk cakupan geologis dari permintaan yang dilakukan. Untuk mempersempit jumlah kelas lebih jauh, digunakan tipe lokasi level 2 area administratif dari hasil yang dikembalikan dari permintaan Google API. Hal ini menyebabkan jumlah kelas turun menjadi 390. Namun, meskipun demikian, tidak ada perbedaan yang signifikan dalam hasil dari Tabel 2. Setelah memeriksa output, dicatat bahwa karena daerah level 2 administrasi daerah berada di tingkat propinsi, cakupannya terlalu luas untuk hanya menggunakan judul untuk mengklasifikasikan postingan.

3.2.3. Simpulan Berdasarkan Klasifikasi Lokasi

Tabel 3. Hasil training data sejumlah ≈1600 dari 8127 sample data

Classifier	Benar	Salah
Naïve Bayes	5130	2997
Td-Tdf	0	8127

Sehubungan dengan klasifikasi berdasarkan lokasi, kurang pas untuk melakukan pengklasifikasi data di Cragslist untuk mencari tempat kost. Namun, perlu upaya untuk melihat apakah Naïve Bayes atau Tf-Idf dapat mengambil berbagai jenis pola yang manusia tidak bisa. Dengan memberikan lebih banyak data pelatihan untuk menemukan beberapa jenis pola. Satu set kelas yang lebih kecil kemungkinan akan sangat bermanfaat juga. Karena semua pembersihan data otomatis, jumlah kelas tidak berhasil dikurangi menjadi hanya jumlah negara seperti yang diharapkan. Mengedit data secara manual yang diambil dari internet sepertinya satu-satunya tindakan yang akan menghasilkan berkurangnya kelas yang mungkin bekerja dengan baik. Namun, mengingat bahwa teks judul untuk iklan sangat kecil dan sering diulang di antara lokasi yang berbeda, sepertinya tidak mungkin untuk mengklasifikasikan lokasi tempat kost dengan mempostingnya di craigslist.

3.3.KLASIFIKASI BERDASARKAN HARGA

3.3.1. Pembatasan Harga

Untuk mengklasifikasikan postingan dari judul ke kisaran harga, pertama-tama perlu menentukan kisaran harga itu sendiri. Untuk percobaan ini, pertama-tama dikumpulkan 8127 sampel tupel judul dan harga, untuk menentukan braket harga sederhana tinggi atau rendah, rata-rata semua harga dihitung dan nilai seratus digunakan sebagai titik cutoff untuk tinggi atau rendah . Rata-rata ini berakhir menjadi ≈1030. Tabel 3 menunjukkan hasil pelatihan pada seperlima data dan memvalidasi sisanya.

3.3.2. Pembersihan Data

Data harga tidak mengalami masalah yang sama dengan data lokasi, di mana memiliki terlalu banyak kelas. Sebaliknya, data harga mengalami masalah distribusi. Sementara rata-rata harga tetap mendekati 1.000.000, ini karena sebagian besar harga tempat kost antara 700.000-1.500.000 Meskipun dapat menganggap rata-rata akan lebih tinggi karena ini, dengan menarik data tidak hanya dari tempat kost tetapi juga

menyewakan iklan, lebih condong dataset ke bagian bawah. Untuk memperbaiki kedua strategi ini dicoba. Duplikat sampel kelas yang kurang terwakili, dan mengumpulkan lebih banyak data dengan harapan menemukan tempat kost yang lebih mahal yang akan membantu distribusi menjadi lebih seragam.

Tabel 4. Hasil traning data ≈2.000 dari 11.124 sample data

Classifier	Benar	Salah
Naïve Bayes	5994	5130
Tf-Tdf	5130	5994

Menggandakan data mudah dilakukan tanpa benar-benar memodifikasi data yang diambil dari Craigslist. Karena distribusi tinggi dan rendah sekitar 63% hingga 37%, dengan hanya menduplikasi sampel kelas tinggi, dataset menjadi 6404 : 7510. Sekali lagi, melatih seperlima dari data ini dan kemudian memvalidasi sisanya membuat para pengklasifikasi selalu menebak satu kelas, mana yang paling sering. Seperti yang terlihat pada Tabel 4, Naïve Bayes selalu memilih kelas yang paling sering digunakan sebagai pilihan default dan Tf-Tdf menggunakan pilihan default apa pun yang ditentukan saat memanggil metode klasifikasi. Menambah lebih banyak posting craigslist yang memiliki biaya lebih tinggi semudah mengarahkan crawler ke bagian tempat kost, kemudian hanya menulis daftar mahal ke file data yang digunakan selama klasifikasi.

Dengan melakukan ini, akan meningkatkan ukuran sampel hingga 14911 dengan 7504 kelas tinggi dan 7407 kelas rendah. Pelatihan berjalan seperti biasa, dengan seperlima dari data pelatihan digunakan. Anehnya, Naïve Bayes berkinerja lebih buruk daripada Tf-Tdf selama menjalankan ini, karena secara konsisten menebak kelas tinggi tetapi pelatihan validasi yang ditetapkan hanya memiliki 5.912 sampel seperti itu. Menambahkan lebih banyak data ke dalam campuran memang membantu classifier beralih dari selalu menebak rendah menjadi selalu menebak rendah. Tetapi tidak memiliki hasil yang diinginkan untuk memberikan lebih banyak informasi judul untuk benar-benar dilatih.

3.3.3. Kesimpulan Klasifikasi Berdasarkan Harga

Seperti halnya klasifikasi lokasi, sepertinya tidak cukup data yang berharga dalam judul posting Craigslist untuk mengklasifikasikan kisaran harga tempat kost dengan benar. Hasil Tf-Tdf mengalami kesulitan mungkin akibat dari ini, karena Tf-Tdf pada dasarnya menentukan arti kata, mungkin tidak ada cukup kata-kata umum antara posting untuk benar-benar menentukan nilai daftar. Ini juga berlaku untuk Naïve Bayes karena variasi teks yang besar menghasilkan probabilitas yang agak rendah begitu bukti ditentukan untuk judul yang diberikan. Mungkin ada terlalu banyak data. Namun, baik data itu sendiri terlalu bias untuk membantu atau sekali lagi, keragaman kata-kata yang berbeda terlalu luas untuk metode penghitungan frekuensi untuk benar-benar membantu dalam klasifikasi.

3.4. PENELUSURAN LEBIH MENDALAM

3.4.1. Akuisisi Data

Karena judul untuk setiap posting sama sekali tidak menyediakan data yang cukup, setelah hasil bagian IV dan V dikumpulkan, alih-alih hanya merayapi halaman indeks yang hanya berisi judul informasi, perayapan baru dibuat yang menggunakan tautan dari halaman tingkat atas untuk mengambil data deskripsi iklan yang lebih panjang. Ini meningkatkan waktu penjelajahan, tetapi data yang dihasilkan lebih deskriptif, dan secara intuitif akan meningkatkan kemampuan kedua pengklasifikasi untuk

menentukan kisaran harga apa yang ada di setiap daftar.

Seperti disebutkan sebelumnya, data pada halaman Craigslist sangat terstruktur, dan memperoleh tautan dari teks halaman tidak sulit. Satu ungkapan reguler sederhana dapat menemukan semua tautan yang mungkin pada halaman dan mengembalikan hasilnya. Dengan mengintegrasikan pencocokan pola ini ke dalam proses pengumpulan judul dan harga, setiap objek iklan dibangun sekaligus dan dapat disimpan untuk digunakan nanti. Sayangnya, seperti disebutkan sebelumnya, tahap akuisisi data percobaan meningkat secara signifikan.

Dari sekadar merayapi 100 atau lebih halaman html menjadi 10.000, peningkatan kali ini memperlambat penelitian secara signifikan - dan terkadang mengakibatkan koneksi disetel ulang oleh server di Craigslist. Selain memperoleh lebih banyak teks per sampel untuk dikerjakan, selama pra-pemrosesan data, stopword tertentu dihapus. Kata berhenti adalah kata-kata yang umum yang tidak menambahkan makna kontekstual yang berharga untuk sampel. Contoh kata-kata tersebut adalah, di, atau, dan. Dengan menghapus kata-kata ini, akan meningkatkan perbedaan antara sampel yang mungkin dengan harapan bahwa kata-kata spesifik yang menunjukkan kisaran harga tertentu akan naik ke atas dan pengklasifikasi akan dapat mengambil pada frase kunci ini.

Tabel 5. Hasil Klasiikasi Data

Classifier	Benar	Salah	Kata Berhenti
Naïve Bayes	1067	667	Dihapus
Naïve Bayes	1067	667	Tertinggal
Tf-Idf	667	1067	Dihapus
Tf-Idf	667	1067	Tertinggal

3.4.2. Hasil Penelusuran Lebih Mendalam

Dengan ukuran sampel 2601 total sampel, 1000 termasuk dalam kisaran harga tinggi, dan 1601 termasuk dalam kisaran harga lebih rendah. Setiap classifier diberi dataset dengan dan tanpa kata-kata berhenti dihapus, Kemudian dilatih pada sepertiga dari kumpulan data dan diberi sampel yang tersisa untuk divalidasi. Sayangnya, dengan kata-kata berhenti atau tidak, tampaknya data itu terlalu banyak untuk ditangani. Seperti upaya klasifikasi sebelumnya, Tf-Idf hanya mengembalikan kelas default, dan Naïve Bayes menebak kelas mana yang paling banyak memberikan contoh.

Dicatat sebelumnya, sementara ini adalah strategi yang baik ketika berhadapan dengan sulit untuk mengklasifikasikan data, itu menunjukkan bahwa Naïve Bayes tidak bisa menangani keragaman besar ruang pencarian. Dengan memasukkan begitu banyak kata ke dalam analisisnya dan begitu banyak variabel, ditemukan bahwa sulit bagi classifier untuk mengembalikan apa pun kecuali kelas default. Tampaknya kedua pengklasifikasi mengalami kesulitan, dan secara intuitif masuk akal bahwa volume ruang pencarian akan sangat besar untuk sesuatu yang sederhana seperti penghitungan frekuensi.

4. KESIMPULAN

Naive Bayes dan Tf-Idf tidak berkinerja baik pada data yang noisy tempat kost dan daftar kamar-untuk-sewa. Data itu sendiri terlalu sulit untuk ditangani oleh pengklasifikasi sederhana; mungkin juga tidak ada korelasi antara apa yang ada dalam judul dan deskripsi iklan dan harganya. Sementara diasumsikan bahwa korelasi itu ada karena manusia dapat melihat teks dan gambar dan melihat nilai di dalamnya, bukan tidak mungkin komputer dapat melakukan hal yang sama. Performa setiap algoritma

secepat yang diharapkan, namun tidak ada hasil nyata yang ditemukan. Default ke satu tebakan untuk seluruh dataset-sementara meminimalkan kesalahan - tidak menunjukkan perbaikan untuk algoritma yang diberikan lebih atau kurang data pelatihan.

5. DAFTAR PUSTAKA

Artaye, Ketut (2015). Implementation of Naïve Bayes Classification method to Predict graduation time of IBI Darmajaya Scholar.

Felida, Naifiri Novitasari (2017). Prediksi waktu Produksi Mebel menggunakan Metode Naïve Bayes.

Feldman, Ronen, dan Sanger, James. (2007). The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data. New York:Cambridge University Press.

Kusrini, & Luthfi, E. T. (2009). Algoritma Data Mining. Yogyakarta: C.V Andi Offset.

Melita, Ria, Victor Amrizal, Hendra Bayu, Taslimun Dirjam, (2018), Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam), JURNAL TEKNIK INFORMATIKA VOL 11 NO. 2, OKTOBER 2018

Nurjoko (2016). Aplikasi Data Mining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma Apriori Di IBI Darmajaya.

Prasetyo, Eko (2014). Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab.