UNIVERSITAS INDONESIA

LAPORAN PROYEK AKHIR

Pendekatan *Predictive Analytics* dalam Prediksi Demografi Pengguna Twitter di Indonesia terhadap Aspek Tipe Akun, Lokasi, dan Ranah Pekerjaan

Kamila Kaffah
1806191225

Nicolas Henry Wijaya
1806191566

Setyawan Pratama
1806191591

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI SISTEM INFORMASI
DEPOK
JULI 2021

BAB I
Pendahuluan

Di era globalisasi saat ini, sosial media berkembang sangat pesat dan menjadi salah satu kegiatan utama seseorang ketika menggunakan internet. Selain itu, media sosial juga menjadi *platform* yang paling efektif bagi perusahaan untuk mendapatkan data yang esensial untuk inovasi dan pengembangan produk (Gozuacik et al., 2021).

Salah satu topik yang saat ini banyak dibahas adalah masalah demografi pengguna media sosial. Hal ini dikarenakan informasi demografi dapat dimanfaatkan untuk berbagai kepentingan, salah satu contohnya adalah kepentingan bisnis dalam mengembangkan fitur suatu produk. Pendekatan yang banyak digunakan untuk mengkategorikan demografi pengguna media sosial adalah pendekatan prediktif analisis.

Media sosial yang banyak digunakan untuk pendekatan prediktif analisis adalah Twitter. Berdasarkan hasil penelitian PeerReach, diketahui bahwa Indonesia tercatat sebagai negara dengan pengguna Twitter terbanyak ketiga di dunia, dengan jumlah 6,5% setelah Amerika Serikat (24,3%), dan Jepang (9,3%). Selain itu, Twitter juga dipandang sebagai media sosial yang dapat digunakan untuk memperoleh data besar dengan biaya yang murah.

Paper penelitian ini akan membahas beberapa pendekatan prediktif analisis terhadap susunan demografi pengguna media sosial. Lalu, penelitian ini terbatas khusus pengguna media sosial Twitter di Indonesia. Terdapat tiga aspek demografi yang akan diprediksi pada penelitian ini, yaitu tipe akun, lokasi akun, dan ranah pekerjaan akun. Selain itu, penelitian ini juga akan membahas pendekatan yang digunakan pada penelitian ini serta perbandingan penggunaan algoritma pada pendekatan *machine learning*.

Diharapkan penelitian ini dapat memberikan gambaran untuk peneliti di masa yang akan datang terkait penggunaan data Twitter untuk analisis prediktif menggunakan *machine learning* dari berbagai pendekatan dan algoritma. Selain itu, penelitian ini dapat memberikan informasi terkait kekurangan dan kelebihan dari beberapa pendekatan dan algoritma *machine learning* yang populer. Dengan begitu, peneliti lain dapat mempersiapkan hal-hal yang sekiranya perlu disiapkan ketika melakukan analisis prediktif, terutama ketika peneliti memilih untuk menggunakan data Twitter.

Struktur penelitian ini terdiri dari pendahuluan, tinjauan pustaka yang

membahas teori terkait penelitian, dilanjuti dengan metodologi, pendekatan penelitian analisis prediktif, dan evaluasi.

BAB II
Tinjauan Pustaka

Bab ini akan menjelaskan tinjauan pustaka yang dilakukan oleh peneliti. Bab ini terdiri dari pembahasan tentang Twitter, *machine learning,* dan *text mining.*

2.1 Twitter

Twitter adalah situs *microblogging* dan saat ini rata-rata pengguna aktif bulanan sejumlah 326 juta, dimana seseorang dapat memposting *updates* singkat yang disebut dengan '*tweets'* ke jaringan seseorang (Moshkovitz & Hayat, 2021). Tweets akan diposting ke *timeline* dan memiliki konteks bahasan yang bervariasi, berkisar dari humor hingga renungan kehidupan, berita terkini dari *outlet* berita, dan pesan langsung dari selebritas, pemerintah, hingga kepala negara. Selain itu, pada saat waktu tertentu, misal ketika krisis bencana alam atau krisis pemerintahan, Tweets juga bisa berisi *update* terbaru dari masyarakat secara *real time.* Berdasarkan hal tersebut, Twitter telah menjadi *platform* untuk berbagi informasi substantif dan bukan sekedar pengalaman sehari-hari (Moshkovitz & Hayat, 2021).

2.2 Machine Learning

Machine learning merupakan serangkaian teknik yang dapat membantu dalam menangani dan memprediksi data yang sangat besar dengan cara mem-presentasikan data-data tersebut dengan algoritma pembelajaran (Danukusumo, 2017).

Berdasarkan pembelajaran dalam *machine learning*, terdapat 3 jenis skenario:

1. *Supervised Learning*

Pembelajaran menggunakan data *training* yang telah dilabeli.

2. *Unsupervised Learning*

Pembelajaran menggunakan data *training* yang tidak diberikan label.

3. *Reinforcement Learning*

Pembelajaran menggunakan campuran data *training* dan *testing* untuk mengumpulkan informasi secara aktif. Metode pembelajaran ini akan menghasilkan *reward* seperti bentuk poin jika model yang dihasilkan semakin baik.

2.3 Text Mining

Text mining merupakan salah satu bagian dari teknik *data mining*. *Text mining* merupakan sebuah proses untuk memperoleh informasi dari sekumpulan teks. *Text mining* juga disebut dengan *text analytics,* yaitu teknik *artificial intelligence* yang mengubah data tidak terstruktur menjadi data terstruktur dengan menggunakan *natural language*

processing (NLP) untuk meningkatkan analisis menggunakan algoritma *machine learning. Text mining* adalah teknik populer di antara *computer science, information science, mathematics,* dan *management fields* untuk *mining intelligence* dari data besar (Kumar et al., 2021).

BAB III
Metodologi

Bab ini akan menjelaskan metodologi yang dilakukan oleh peneliti. Bab ini terdiri dari tiga sub bab, yaitu metodologi penelitian, data, dan *tools* yang digunakan.

3.1 Metodologi Penelitian

Dalam mengerjakan penelitian ini, tim peneliti melakukan setidaknya 7 tahapan secara berurutan. Setiap tahapan dilakukan guna mendapatkan hasil atau *output* tertentu yang nantinya hasil tersebut digunakan pada tahapan selanjutnya. Bagian dibawah ini menjelaskan tahapan-tahapan yang peneliti lakukan.

3.1.1 Membaca riset

Menurut Djunaedi (2000), seseorang perlu mengetahui, mengenal dan memahami penelitian-penelitian terdahulu yang serupa dengan topik yang diteliti agar peneliti dapat mempertanggungjawabkan cara yang digunakan dalam meneliti permasalahan yang dihadapi. Dalam hal ini, peneliti juga melakukan hal yang sama yaitu dengan membaca studi-studi terdahulu yang berkaitan dengan klasifikasi akun Twitter berdasarkan profil pengguna dan distribusi pengikut. Proses ini penulis laksanakan dengan tujuan meningkatkan pemahaman teoritis dan menambah referensi model serta pendekatan penelitian yang sesuai.

3.1.2 Mengumpulkan Data

Pada penelitian ini digunakan 2 jenis data, yaitu data *training* dan data *testing*. Kedua jenis data ini merupakan data yang berasal dari Twitter. Data training digunakan untuk melatih model *machine learning* yang peneliti buat. Sedangkan data *testing* digunakan untuk memprediksi beberapa atribut sesuai dengan permintaan soal. Untuk mengumpulkan data *training*, peserta dan asisten dosen mata kuliah Analitika Media Sosial melakukan pengumpulan data *username* akun Twitter secara kolektif. Akun Twitter yang berhasil dikumpulkan dari pengumpulan kolektif ini adalah sebanyak 5213 *username*. Sedangkan untuk data *testing* sudah disediakan oleh tim dosen dan asisten berupa kumpulan *username* Twitter.

3.1.3 Melakukan anotasi data

Data *training* yang telah dikumpulkan oleh peserta dan asisten dosen mata kuliah Analitika Media Sosial tidak semuanya memiliki label kategori isaPerson, location

dan jobArea. Karena kurangnya informasi yang memadai untuk melakukan proses *machine learning* menggunakan *supervised learning*, peneliti memutuskan untuk melakukan anotasi data-data yang belum lengkap.

3.1.4 Data *Preprocessing*

Data yang tidak berkualitas akan menghasilkan prediction atau *decision* yang buruk. Untuk mendapatkan modelling yang baik, perlu dilakukan data *preprocessing* terlebih dahulu. Data *preprocessing* dilakukan dengan tahapan pembersihan data, penghapusan atribut yang tidak dibutuhkan dan konversi data. Dari proses data *preprocessing*, terdapat 1507 baris data yang di-*drop* oleh peneliti sehingga menyisakan 3706 baris yang digunakan untuk data *training*.

3.1.5 Menentukan Metode yang Akan Digunakan

Terdapat dua pendekatan yang digunakan oleh peneliti untuk memprediksi label demografi, yaitu pendekatan *supervised learning* dan *rule based.* Secara umum, peneliti menggunakan kombinasi kedua pendekatan tersebut untuk menyelesaikan suatu model untuk memprediksi label demografi.

3.1.6. Membangun Model

Peneliti membangun model berdasarkan pendekatan metode yang sudah ditentukan sebelumnya. Penggunaan jumlah data dan fitur disesuaikan oleh peneliti sesuai dengan kebutuhan untuk masing-masing model.

3.1.7. Submisi

Setelah membangun model, peneliti melakukan submisi ke *grader* yang telah disediakan oleh asisten dosen. *Grader* tersebut dapat diakses pada halaman http://anamedsos2021.herokuapp.com/.

3.1.8. Evaluasi

Hasil perhitungan yang dikeluarkan oleh *grader* terdiri dari *accuracy, precision, recall,* dan *F1-score.* Selain itu, terdapat pula detail perhitungan yang disediakan oleh *grader*, yaitu perhitungan *macro-average, micro-average,* dan *weighted-average.* Terakhir, *grader* menyediakan informasi *confusion matrix* yang dapat dimanfaatkan untuk melihat jumlah data yang berhasil diprediksi pada setiap kategorinya.

Hasil perhitungan *grader* yang dijelaskan pada paragraf sebelumnya digunakan penulis untuk melakukan evaluasi. Setelah itu, peneliti melakukan perbaikan kode berdasarkan hasil evaluasi dan diskusi antar peneliti.

3.2 Data

Seperti yang dijelaskan pada bagian sebelumnya, penelitian ini menggunakan 2 jenis data, yaitu data *training* dan data

testing. Setelah kedua data tersebut dikumpulkan *username*-nya, peneliti mengumpulkan beberapa atribut lain dari *username* yang bersangkut. Tujuannya adalah agar atribut-atribut tersebut dapat dijadikan fitur dalam model *machine learning* yang peneliti buat sehingga dapat membantu prediksi menjadi lebih akurat. Tabel dibawah ini merupakan nama dan tipe data atribut yang peneliti kumpulkan

Nama Atribut	Deskripsi
default_profile	Menunjukkan apakah akun Twitter terkait menggunakan tema atau latar belakang bawaan dari Twitter
default_profile_image	Menunjukkan apakah akun Twitter tersebut menggunakan foto profil bawaan dari Twitter.
description	Menunjukkan isi dari kolom *bio* akun Twitter tersebut.
favourites_count	Menunjukkan jumlah *tweet* yang difavoritkan oleh akun Twitter tersebut.
followers_count	Menunjukkan jumlah *follower* dari akun Twitter tersebut.
friends_count	Menunjukkan jumlah *following* dari akun Twitter tersebut.
geo_enabled	Menunjukkan apakah akun tersebut mencantumkan lokasi pada profilnya.
profile_location	Menunjukkan lokasi dari akun Twitter tersebut yang terdapat pada profil.
name	Menunjukkan nama dari akun Twitter tersebut.
protected	Menunjukkan apakah akun Twitter tersebut membatasi privasi akunnya.
statuses_count	Menunjukkan jumlah *tweet* yang dibuat oleh akun Twitter tersebut.
url	Menunjukkan *link* yang ada pada profil akun Twitter tersebut.
verified	Menunjukkan apakah akun twitter tersebut termasuk akun yang terverifikasi oleh Twitter.
username	Menunjukkan *string* unik yang mengidentifikasi akun Twitter tersebut.

Pada kenyataannya, tidak semua atribut yang peneliti kumpulkan digunakan pada

setiap prediksi yang peneliti lakukan. Atribut yang peneliti gunakan untuk memprediksi disesuaikan dengan *output* prediksi tersebut. Hal ini akan dibahas lebih lanjut pada bagian Pendekatan Eksperimen.

Data atribut-atribut diatas peneliti peroleh dengan menggunakan teknik *crawling*, dimana peneliti Python dan *library* Tweepy untuk membantu proses ini. *Library* Tweepy memungkinkan peneliti untuk terhubung dengan data-data yang ada di media sosial Twitter sehingga peneliti dapat mengumpulkan data-data atribut yang peneliti perlukan. Data-data atribut peneliti ambil berdasarkan *username* dari akun Twitter terkait, kemudian peneliti masukan ke dalam file .csv yang nantinya digunakan untuk memprediksi.

Setelah mengumpulkan semua data atribut, dilakukan proses anotasi dengan melakukan observasi untuk setiap kategori prediksi yaitu isaPerson, location, dan jobArea. isaPerson merupakan atribut dengan tipe data *boolean* yang menunjukkan apakah akun yang diobservasi merupakan akun perseorangan atau akun yang bukan orang. location merupakan atribut yang mengelompokkan akun pada 9 kelompok lokasi. jobArea merupakan atribut yang mengelompokkan akun pada 11 kelompok pekerjaan.

3.3 *Tools* yang Digunakan

Pada pengerjaan proyek ini, peneliti menggunakan *tools* untuk melakukan *crawling* data dan membangun model. Beberapa *tools* yang peneliti gunakan antara lain:

- Tweepy. *Library* ini digunakan untuk *crawling* data dari media sosial Twitter.
- Pandas. *Library* ini digunakan untuk mengolah data yang ada pada dokumen .csv.
- Numpy. *Library* ini digunakan untuk membantu mengolah *object ndarray.*
- Sklearn. *Library* ini digunakan untuk membantu implementasi *machine learning.*

BAB IV
Pendekatan Eksperimen

Penelitian ini akan menggunakan pendekatan analisis prediktif terhadap tiga atribut, yaitu tipe akun, lokasi, dan pekerjaan.

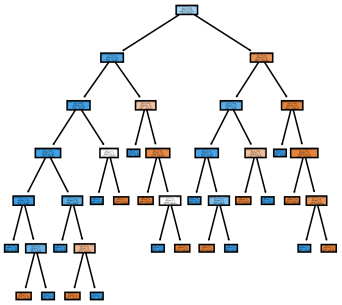
4.1 Prediksi tipe akun

Pada prediksi isaPerson, masing-masing akun Twitter akan diklasifikasikan menjadi 2 kelompok yaitu person dan non person. Akun person adalah akun personal yang dimiliki oleh seorang individu (ditulis dengan kategori angka 1). Akun non person adalah akun yang bukan dimiliki oleh individu, dapat berupa komunitas,

lembaga, bisnis (ditulis dengan kategori angka 0). Untuk memprediksikan atribut tipe akun, peneliti menggunakan data *structured* dari profil yaitu favourites_count, default_profile_image, protected dan friends_count.

Peneliti juga melakukan pendekatan rule-based menggunakan description dan username pada profil. Pendekatan ini dilakukan dengan mengklasifikasikan akun yang memiliki kata "official", "akun resmi" dan "akun twitter resmi" pada username atau profil menjadi akun *non person*.

Algoritma yang digunakan untuk melakukan prediksi ini adalah *random forest classifier*. Algoritma *random forest classifier* adalah salah satu algoritma *machine learning* untuk permasalahan klasifikasi. Secara garis besar *random forest classifier* merupakan sekumpulan *decision tree*, dimana *decision tree* sendiri juga merupakan algoritma *machine learning* yang sekumpulan keputusan yang berurutan dari paling atas hingga mencapai keputusan terakhir yang disusun seperti pohon. Keputusan terakhir inilah yang nantinya menjadi hasil prediksi dari model *machine learning* yang dibuat. Algoritma *random forest classifier* relatif lebih kompleks daripada algoritma *decision tree* karena *random forest classifier* sendiri terdiri dari banyak *decision tree*. Gambar dibawah ini menjadi ilustrasi bagaimana algoritma *random forest* bekerja.



Algoritma ini dipilih karena merupakan algoritma yang memberikan nilai average tertinggi (81.25%) saat dicoba dengan algoritma-algoritma lainnya pada fase *training*.

4.2 Prediksi lokasi

Pada penelitian ini, prediksi untuk atribut lokasi dibagi menjadi 9 kategori. Prediksi ini dilakukan untuk akun yang bertipe *person* maupun *not person.* Berikut adalah daftar kategori yang akan diprediksi untuk setiap akun.

- sumatera
- jabodetabek
- jawa barat dan banten (selain bogor)
- jawa tengah dan yogyakarta
- jawa timur
- kalimantan
- sulawesi
- bali dan nusa tenggara
- maluku dan papua

Pendekatan yang digunakan untuk memprediksi lokasi dari suatu akun adalah

pendekatan *rule base.* Untuk dapat memprediksi lokasi, peneliti menggunakan dua fitur dari *testing data*, yaitu fitur profile_location dan fitur description. Sebagai langkah awal, peneliti membuat sebuah leksikon atau kamus yang berisi kumpulan kata yang bersesuaian dengan 9 kategori lokasi yang ada. Kata-kata yang dimasukan ke dalam setiap kategori lokasi dipilih berdasarkan daerah-daerah yang termasuk ke dalam kategori tersebut. Sebagai contoh, untuk kategori Sumatera, peneliti memasukkan kata berupa provinsi-provinsi yang ada di pulau Sumatera, yaitu Sumatera Barat, Sumatera Selatan, dan seterusnya. Kemudian kamu juga memasukkan daerah-daerah yang lebih kecil seperti ibukota, kabupaten, hingga kelurahan.

Selanjutnya, peneliti akan memeriksa fitur profile_location untuk setiap data pengguna dan melakukan *mapping* nilai profile_location tersebut dengan leksikon yang sudah dibuat. Apabila fitur profile_location pada data tidak memiliki nilai, peneliti akan melakukan langkah berikutnya, yaitu menggunakan fitur description.

Setiap kata pada fitur description akan diambil dan akan dilakukan *mapping* dengan leksikon yang sudah dibuat sebelumnya. Langkah ini efektif untuk menangani kasus data yang memiliki informasi lokasi pada fitur description. Terakhir, apabila terdapat data yang gagal ditangani oleh pemetaan fitur profile_location dan fitur description maka secara otomatis data akan dipetakan ke dalam kategori jabodetabek karena kategori tersebut merupakan modus dari data. Contoh kasus yang gagal ditangani oleh pemetaan fitur profile_location dan fitur description adalah apabila data tidak memuat informasi lokasi pada kedua fitur tersebut.

4.3 Prediksi pekerjaan

Pada penelitian ini, prediksi untuk atribut jobArea dibagi menjadi sebelas kategori. Berikut adalah daftar kategori yang akan diprediksi untuk setiap akun.

- pendidikan dan penelitian
- sains dan teknologi
- kesehatan
- ekonomi dan bisnis
- sosial kemasyarakatan
- media
- hospitality & tourism
- olahraga
- hiburan
- seni
- hobi

Untuk memprediksi jobArea dari suatu akun, peneliti melakukan 2 bentuk pendekatan. Pertama adalah dengan menggunakan leksikon jobArea. Leksikon ini peneliti buat dengan menggunakan

ekstraksi unigram dan bigram. Peneliti menggunakan CountVectorizer yang merupakan *library* bawaan dari sklearn untuk mendapatkan 2000 fitur dengan frekuensi kemunculan paling banyak dari *description* profil. Kata-kata yang didapatkan kemudian dipetakan secara manual ke dalam kategori pekerjaan yang sesuai.

Pendekatan yang kedua adalah dengan *supervised-based learning* menggunakan algoritma *random forest classifier*. Pada proses *training*, tidak semua data training yang dikumpulkan dipakai. Peneliti menyeleksi kembali record-record yang digunakan sebagai data training berdasarkan ketersediaan gold label *feature* jobArea. Data training yang digunakan untuk membangun model ini adalah sebanyak 1250 data. Pendekatan kedua merupakan alternatif dari pendekatan pertama apabila description profile akun Twitter yang diprediksi kosong.

BAB V
Evaluasi

Pada bagian ini, peneliti melihat dan membandingkan hasil prediksi dengan data aktualnya. Evaluasi dilakukan dengan beberapa matriks, seperti *a*ccuracy, *precision*, *recall*, dan F1-*score*. Dengan dilakukannya proses evaluasi, peneliti berharap mengetahui kesalahan dan kekurangan dari pendekatan yang dilakukan sehingga dapat membuat solusi yang lebih baik lagi.

5.1 Pengolahan data

Pada proses pengolahan data, peneliti melakukan sejumlah pekerjaan seperti pengisian nilai yang kosong dengan 0 atau ‘’ serta melakukan normalisasi kolom yang bernilai integer seperti followers_count, statuses_count, favourites_count, dan kolom lain yang biasanya diiringi oleh _count. Evaluasi dalam proses pengolahan data adalah pada normalisasi. Setelah proses normalisasi, peneliti mengubah tipe data kolom menjadi integer32 sehingga diduga mengurangi esensi dari data dimana saat ini data dari kolom hanya bernilai 0 atau 1.

Kedepannya, peneliti dapat mencoba untuk mempertahankan tipe data float32 sehingga model dapat lebih akurat karena mendapatkan data yang kontinu dibanding dengan data boolean.

5.2 Prediksi tipe akun

Prediksi dengan peringkat paling tinggi dicapai dengan menggunakan pendekatan structured menggunakan fitur protected, favourites_count, default_profile_image, friends_count menggunakan algoritma *random forest classifier* dan pendekatan *rule-based* untuk deskripsi dan username akun. Hasil yang didapatkan dari

submission ini adalah 83.41% accuracy, 84.45% precision, 69.90% recall dan 73.17% F1.

Peneliti melakukan evaluasi pada model untuk mengetahui kekurangan yang ada dalam model, berikut merupakan 10 record pertama data false positive.

Protected	Favourite	DPI	Friend
F	0	F	8
F	344	F	221
F	1678	F	10
F	29	F	45415
F	29	F	203
F	402	F	250
F	1235	F	3
F	15	F	122
F	470	F	1209
F	0	F	511
F	59	F	251

Berdasarkan analisa peneliti terhadap data-data yang masuk sebagai false positive, ditemukan bahwa model gagal melakukan prediksi untuk akun-akun dengan jumlah friends_count yang tinggi. Dari data, ditemukan bahwa rata-rata favourites_count dan friends_count dari akun tipe non person adalah 1207 dan 639 sedangkan rata-rata akun person adalah 2129 dan 925. Hal ini sesuai dengan cara model melakukan klasifikasi yaitu akun person memiliki favourites_count atau friends_count yang tinggi.

Namun, cukup banyak akun Twitter bisnis dengan tingkat interaksi yang tinggi dan akun Twitter seperti komunitas dan menfess yang memiliki jumlah friends_count yang tinggi. Untuk fitur protected dan default_profile_image tidak ditemukan pengaruh yang tinggi karena hanya ditemukan 2 akun protected dan 3 akun dengan default profile image pada kategori ini.

Selain false positive, peneliti juga berusaha mengetahui kekurangan lain dalam model (false negative), berikut merupakan 10 record pertama data false negative.

Protected	Favourite	DPI	Friend
F	2	F	91
F	50	F	324
F	3172	F	3119
F	263	F	719
F	9807	F	514
F	26882	F	1438
F	77	F	948
F	5	F	20
F	244	F	413
F	1897	F	351

Berdasarkan analisa peneliti terhadap data-data yang masuk sebagai false

negative, kebanyakan merupakan individu yang “terkenal” dapat berupa tokoh masyarakat atau influencer. Terdapat juga akun-akun Twitter yang memang kurang interaktif dengan akun lainnya. Tingkat favourites_count dan friends_count yang rendah menyebabkan akun-akun tersebut dikelompokkan sebagai akun bisnis. Faktor lain yang menyebabkan error false positive adalah penggunaan “official”, “akun resmi”, “akun twitter” oleh influencer atau tokoh masyarakat yang langsung peneliti golongkan sebagai akun non person.

Kedepannya, model ini dapat ditingkatkan dengan data-data lain yang bisa didapatkan seperti jumlah postingan yang mengandung mention ke pengguna lain, jumlah tweet yang dikutip, jumlah tweet dalam kaitannya dengan tempat (menggunakan lexicon lokasi, jumlah tweet yang menggunakan fitur geotag, interval waktu tweet. Dengan data-data seperti ini, model dapat lebih baik lagi dalam menentukan akun person atau non person walaupun tentunya disertai drawback yaitu proses pengolahan data yang lebih sulit dan lambat.

5.3 Prediksi lokasi

Peneliti berhasil mendapatkan nilai evaluasi dengan detail: *accuracy* sebesar 57,13; *precision* sebesar 85,62; *recall* sebesar 54.26; F1-*score* 62,22. Nilai tersebut peneliti dapatkan setelah beberapa kali melakukan submisi. Untuk mendapatkan nilai evaluasi yang lebih baik, peneliti secara kolektif menambah jumlah kata yang ada pada leksikon lokasi peneliti. Kata-kata yang ditambahkan adalah daerah atau lokasi yang lebih kecil ruang lingkupnya daripada lokasi-lokasi yang sudah ada. Peneliti juga menambahkan beberapa lokasi atau daerah besar yang terlewat saat mendaftarkan leksikon untuk pertama kalinya. Dengan leksikon yang lebih lengkap, akhirnya peneliti bisa mencapai nilai evaluasi dengan rata-rata 64,8.

Untuk melakukan evaluasi, peneliti membandingkan hasil prediksi dengan data aktual yang disediakan oleh tim asisten dosen. Karena peneliti menggunakan teknik *rule-based* untuk mengerjakan persoalan ini, evaluasi dilakukan dengan cara melihat berapa banyak data yang berhasil diprediksi dan berapa banyak data yang gagal diprediksi. Tabel dibawah ini menunjukkan distribusi data aktual yang kami prediksi.

Kategori	**Persentase distribusi data**
jabodetabek	25,61%
sumatera	19,35%
jawa barat dan banten	14,86%
jawa tengah dan yogyakarta	13,57%

jawa timur	10,57%
maluku dan papua	4,36%
kalimantan	4,18%
bali dan nusa tenggara	3,75%
sulawesi	3,75%

Dari tabel tersebut, dapat diketahui jika distribusi paling banyak adalah pada kategori jabodetabek. Kemudian, tabel dibawah ini menunjukkan berapa persentase keberhasilan kami dalam memprediksi lokasi dari sekumpulan akun Twitter yang diberikan.

Kategori	**Persentase berhasil diprediksi**
jabodetabek	97,12%
sumatera	35,87%
jawa barat dan banten	38,02%
jawa tengah dan yogyakarta	44,34%
jawa timur	37,79%
maluku dan papua	73,24%
kalimantan	57,35%
bali dan nusa tenggara	47,54%
sulawesi	55,74%

Berdasarkan tabel tersebut dapat diketahui bahwa peneliti berhasil memperoleh prediksi paling akurat untuk kategori jabodetabek, yaitu sebesar 97,12%. Sedangkan prediksi paling buruk adalah pada kategori sumatera, yaitu hanya sebesar 35,87%. Jumlah data yang diprediksi dengan benar adalah sebesar 42,82% data.

Salah satu faktor mengapa kategori jabodetabek mendapatkan persentase yang tinggi adalah karena tim peneliti membuat sebuah aturan dimana jika atribut profile_location dan description tidak mengandung satupun kata yang ada pada leksikon, maka secara otomatis lokasi dari akun Twitter tersebut masuk ke dalam kategori jabodetabek. Aturan ini dibuat dengan dasar wilayah jabodetabek merupakan wilayah yang memiliki jumlah penduduk terbanyak, sehingga besar kemungkinan jika ada banyak akun Twitter yang masuk ke dalam kategori ini.

Prediksi yang peneliti buat bisa dibilang belum cukup baik karena prediksi yang benar bahkan belum mencapai setengah dari total data yang ada. Peneliti menyadari perlu adanya improvisasi dari pendekatan yang sudah dilakukan atau bahkan memilih pendekatan yang lebih baik.

5.4 Prediksi pekerjaan

Prediksi dengan peringkat paling tinggi dicapai dengan menggunakan leksikon dan pendekatan *structured* menggunakan fitur protected, profile_use_background_image, friends_count, default_profile_image, followers_count, menggunakan algoritma *random forest classifier*. Hasil yang

didapatkan dari submission ini adalah 37.65% accuracy, 33.93% precision, 32.73% recall dan 31.56% F1.

Evaluasi peneliti lakukan dengan melihat jumlah data benar untuk setiap kategori yang dibandingkan dengan data yang diberikan oleh asisten dosen.

Kategori	**Persentase berhasil memprediksi**
pendidikan dan penelitian	41,7%
sains dan teknologi	34,5%
kesehatan	42,8%
ekonomi dan bisnis	37,6%
sosial kemasyarakatan	48,8%
media	34,4%
hospitality & tourism	21,4%
olahraga	33,3%
hiburan	35,8%
seni	24,2%
hobi	18,3%

Menurut evaluasi dari peneliti, kategori hobi menjadi kategori paling sulit diprediksi karena tidak terdapat kata-kata yang banyak berafiliasi dengan hobi. Terdapat cukup banyak artis atau tokoh yang memiliki komunitas ditambah dengan komunitas aktivitas lain yang beragam. Kategori hobi juga biasanya overlap dengan kategori lain seperti olahraga, hiburan dan sosial kemasyarakatan sehingga sulit untuk diprediksi.

Hal lain yang menyebabkan buruknya prediksi adalah karena adanya kondisi dimana suatu akun Twitter menulis beberapa bentuk pekerjaan pada description. Pendekatan lain untuk mengetahui pekerjaan saat deskripsi kosong dengan *supervised based learning* juga tidak banyak membantu karena masing-masing pekerjaan tidak memiliki perbedaan yang signifikan jika diperhatikan dari data profil. Akurasi dari *supervised based learning* yang digunakan adalah 3,66%. Dengan akurasi seperti ini, mungkin akan lebih baik jika peneliti menggunakan teknik *random* 11 kategori untuk *non person* dan *random* 10 kategori untuk *person* yang memiliki kemungkinan sebesar 9,1% dan 10%.

Kedepannya, prediksi jobArea dapat dilakukan melalui pendekatan *scraping* dan menggunakan *unstructured* data lain seperti *tweet* untuk memprediksi pekerjaan dari suatu akun Twitter.

Daftar Pustaka

Arifin, BH. (2015). Pengguna Twitter Indonesia Terbanyak Ketiga Dunia. http://www.enciety.co/penggunatwitter-indonesia-terbanyak-ketigadunia/.

Daouadi, K. E., Zghal Rebaï, R., & Amous, I. (2019). Organization, Bot, or Human: Towards an Efficient Twitter User Classification. Computación y Sistemas, 23(2). https://doi.org/10.13053/cys-23-2-3192

Gozuacik, N., Sakar, C., & Ozcan, S. (2021). Social media-based opinion retrieval for product analysis using multi-task deep neural networks. Expert Systems With Applications, 183, 115388. https://doi.org/10.1016/j.eswa.2021.115388

Kumar, S., Kar, A., & Ilavarasan, P. (2021). Applications of text mining in services management: A systematic literature review. *International Journal Of Information Management Data Insights*, *1*(1), 100008. https://doi.org/10.1016/j.jjimei.2021.100008

Moshkovitz, K., & Hayat, T. (2021). The rich get richer: Extroverts' social capital on twitter. *Technology In Society*, *65*, 101551. https://doi.org/10.1016/j.techsoc.2021.101551

Oentaryo, R. J., Low, J.-W., & Lim, E.-P. (2015). Chalk and Cheese in Twitter: Discriminating Personal and Organization Accounts. Lecture Notes in Computer Science, 465–476. https://doi.org/10.1007/978-3-319-16354-3_51

Yan, L., Ma, Q., & Yoshikawa, M. (2013). Classifying Twitter Users Based on User Profile and Followers Distribution. Lecture Notes in Computer Science, 396–403. https://doi.org/10.1007/978-3-642-40285-2_34

Preoţiuc-Pietro, D., Lampos, V., & Aletras, N. (2015). An analysis of the user occupational class through Twitter content. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). https://doi.org/10.3115/v1/p15-1169