TUGAS AKHIR – KS 141501

PROTOTIPE EKSTRAKSI PROFIL DEMOGRAFI PELANGGAN MENGGUNAKAN METODE TEMPLATE FILLING UNTUK PERSONALISASI PENCARIAN PRODUK

PROTOTYPE OF CUSTOMER DEMOGRAPHIC PROFILE EXTRACTION USING TEMPLATE FILLING METHODOLOGY FOR PERSONALIZED PRODUCT SEARCHING

ADI SATRIA
NRP 5213 100 180

Dosen Pembimbing :
Rully Agus Hendrawan, S.Kom., M.Eng.

DEPARTEMEN SISTEM INFORMASI
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember
Surabaya 2018

TUGAS AKHIR – KS 141501

PROTOTIPE EKSTRAKSI PROFIL DEMOGRAFI PELANGGAN MENGGUNAKAN METODE TEMPLATE FILLING UNTUK PERSONALISASI PENCARIAN PRODUK

ADI SATRIA
NRP 5213 100 180

Dosen Pembimbing :
Rully Agus Hendrawan, S.Kom., M.Eng.

DEPARTEMEN SISTEM INFORMASI
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember
Surabaya 2018

FINAL PROJECT – KS 141501

PROTOTYPE OF CUSTOMER DEMOGRAPHIC PROFILE EXTRACTION USING TEMPLATE FILLING METHODOLOGY FOR PERSONALIZED PRODUCT SEARCHING

ADI SATRIA
NRP 5213 100 180

SUPERVISOR:
Rully Agus Hendrawan, S.Kom., M.Eng.

DEPARTMENT OF INFORMATION SYSTEMS
Faculty of Information Technology and Communcations
Institut Teknologi Sepuluh Nopember
Surabaya 2018

LEMBAR PENGESAHAN

PROTOTIPE EKSTRASI PROFIL DEMOGRAFI PELANGGAN MENGGUNAKAN METODE TEMPLATE 'ILLING UNTUK PERSONALISASI PENCARIAN PRODUK

TUGAS AKHIR

Disusun untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember

Oleh :

ADI SATRIA
NRP 5213 100 180

Surabaya, Januari 2018

Plh Kepala
Departemen Sistem Informasi

Edwin Riksakomara, S.Kom, MT.
NIP 196907252003121001

LEMBAR PERSETUJUAN

PROTOTIPE EKSTRASI PROFIL DEMOGRAFI PELANGGAN MENGGUNAKAN METODE TEMPLATE FILLING UNTUK PERSONALISASI PENCARIAN PRODUK

TUGAS AKHIR

Disusun untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember

Oleh :

ADI SATRIA
NRP 5213 100 180

Disetujui Tim Penguji : Tanggal Ujian : 10 Januari 2017
Periode Wisuda : Maret 2017

Rully A. Hendrawan, S.Kom., M.Eng **(Pembimbing I)**

Amna Shifia Nisafani, S.Kom, M.Sc **(Penguji I)**

Mahendrawati Er., S.T., M.Sc., Ph.D **(Penguji II)**

PROTOTIPE EKSTRAKSI PROFIL DEMOGRAFI PELANGGAN MENGGUNAKAN METODE TEMPLATE FILLING UNTUK PERSONALISASI PENCARIAN PRODUK

Nama Mahasiswa : Adi Satria
NRP : 5213 100 180
Jurusan : Sistem Informasi FTIf-ITS
Pembimbing 1 : Rully Agus Hendrawan, S.Kom., M.Eng.

ABSTRAK

Saat ini seluruh situs e-commerce memiliki pelanggan yang bermacam-macam profil demografinya dan masing-masing profil membutuhkan pendekatan yang berbeda pula. Agar e-commerce dapat memiliki hubungan yang baik dengan pelanggannya penting bagi e-commerce untuk mengetahui demografi pelanggannya terlebih dahulu, oleh sebab itu e-commerce perlu melakukan pemetaan dan identifikasi terhadap pelanggannya. Kondisi saat ini upaya dalam melakukan pemetaan dan identifikasi demografi pelanggan masih dilakukan secara manual yaitu dengan menyocokkan data pelanggan yang tersedia dengan atribut demografi secara satu-persatu. Tentu kondisi ini memakan beban kerja dan waktu yang lebih, dan apabila tidak dilakukan maka akan sulit ditentukan strategi hubungan pelanggan yang tepat sebab tidak tersedianya data profil demografi pelanggan.

Dengan demikian diperlukan suatu sistem yang secara otomatis melakukan identifikasi demografi pelanggan berdasarkan ketersediaan data dan menampilkan daftar profil demografi pelanggan secara terstruktur. Pada penelitian ini digunakan metode pembuatan template filling dengan pendekatan finite-state cascade untuk mengambil dan menyimpan informasi demografi pelanggan, dan menampilkan informasi profil demografi secara terstruktur. Luaran dari penelitian ini adalah

perangkat lunak yang melakukan ekstraksi profil demografi pelanggan dan menampilkan informasi demografi pengguna dengan terstruktur.

Kata Kunci: Ekstraksi Informasi, Demografi, Template Filling

PROTOTYPE OF CUSTOMER DEMOGRAPHIC PROFILE EXTRACTION USING TEMPLATE FILLING METHODOLOGY FOR PERSONALIZED PRODUCT SEARCHING

Student Name	**: Adi Satria**
NRP	**: 5213 100 180**
Department	**: Sistem Informasi FTIf-ITS**
Supervisor 1	**: Rully Agus Hendrawan, S.Kom., M.Eng.**

ABSTRACT

Nowadays all e-commerce websites have customers of various demograpic profiles and each profile requires different approach. In order for e-commerce to have good relationship with its customer, it is important for e-commerce to have knowledge about the customer demographic aspects first, therefore e-commerce need to do the mapping and identification to its customer. The current state of the effort to do the mapping and identification of the customer is still done manually by available customer data with demographic attributes one-by-one. This condition leads to more loadwork and time consuming, and if it's not performed then it will be difficult to determine the appropiate strategy regarding customer relationship because of the unavailability of customer demographic profle data.

Thus, a system that automatically identifies customer demographics based on data availability and displays a list of customer demographic profiles in a structured manner is needed. This research used template filling method with finite-state cascade approach to retrieve and store customer demographic information, and display the demographic profile information in a structured manner. The output of this research is software that extracts customer demographic profiles and displays user demographic information in a structured way.

Keywords: ***Information Extraction, Demographic, Template Filling***

KATA PENGANTAR

Puji dan syukur penulis tuturkan ke hadirat Allah SWT, Tuhan Semesta Alam yang telah memberikan kekuatan dan hidayah-Nya kepada penulis sehingga penulis mendapatkan kelancaran dalam menyelesaikan tugas akhir dengan judul:

PROTOTIPE EKSTRAKSI PROFIL DEMOGRAFI PELANGGAN MENGGUNAKAN METODE TEMPLATE FILLING UNTUK PERSONALISASI PENCARIAN PRODUK

yang merupakan salah satu syarat kelulusan pada Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember Surabaya.

Terima kasih penulis sampaikan kepada pihak-pihak yang telah mendukung, memberikan saran, motivasi, semangat, dan bantuan baik berupa materiil maupun moril demi tercapainya tujuan pembuatan tugas akhir ini. Tugas akhir ini tidak akan pernah terwujud tanpa bantuan dan dukungan dari berbagai pihak yang sudah melauangkan waktu, tenaga dan pikirannya. Secara khusus penulis akan menyampaikan ucapan terima kasih yang sebanyak-banyaknya kepada:

1) Orang tua dan saudara penulis, Bapak Ir. Pandu Angklasito, Ibu Drg. Hegar Ningrum, Bangkit Oetomo, dan Dian Fitri yang telah memberikan motivasi, semangat, keyakinan, kasih sayang serta doa sehingga penulis mampu menyelesaikan pendidikan S1 ini dengan baik.
2) Bapak Rully Agus Hendrawan, S.Kom., M.Eng. selaku dosen pembimbing yang telah dengan sabar dan telaten memberikan ilmu, petunjuk, dan motivasi sehingga penulis dapat menyelesaikan Tugas Akhir ini.
3) Pri Rezki, Umar Al Aqsho yang telah membantu merancang sistem pengambilan data dan Natanael Yabes Wirawan yang telah menyediakan sistem penyedia data.

4) Bapak Faizal Johan Atletiko S.Kom., M.T. selaku dosen wali penulis selama menempuh pendidikan di Jurusan Sistem Informasi yang telah memberikan pengalaman serta nasehat kepada penulis selama ini.

5) Ibu Mahendrawathi ER, ST., MSc., Ph.D., Ibu Amna Shifia Nisafani, S.Kom, M.Sc. selaku dosen penguji yang telah memberikan kritik saran, dan masukan yang berharga sehingga dapat menyempurnakan Tugas Akhir ini.

6) Seluruh dosen pengajar beserta staf dan karyawan di Jurusan Sistem Informasi, Fakultas Teknologi Informasi ITS Surabaya yang telah memberikan ilmu dan pengalaman yang berharga kepada penulis selama ini.

7) Rekan-rekan mahasiswa Jurusan Sistem Informasi BELTRANIS serta anggota Lab SE atas semua bantuan ketika penulis kuliah di Sistem Informasi.

8) Teman-teman veteran dalam menyelesaikan tugas akhir Naufal Aditya, Alvin Darari, Robbigh Faubendri, Yeremia Aha, Rizki Eka, Aditya Adrevi, Imam Hadi, Adham Adi, Ihsan Pradipta, Nugraha Prihardika yang terus menyemangati penulis dalam pengerjaan tugas akhir.

9) Pasukan Adit United yang turut menemani dan mendukung penulis: Adit, Icam, Lugas, Jisung, Ucan, Adam, Wira, Inyong, serta teman-teman pasukan Adit United lainnya yang penulis belum dapat sebutkan satu-persatu.

10) Pasukan Idan United yang turut menemani dan mendukung penulis: Idan, Rosu, Umar, Haikal, Om Bimo, serta teman-teman pasukan Idan United yang penulis belum dapat sebutkan satu-persatu.

11) Pasukan Kentang United yang turut menemani dan mendukung penulis: Adit Kentang, Pakaya, Bia, Imam.

12) Pak Gendut, Bapak Takur, dan Bapak Tembong yang telah meramaikan kehidupan perkuliahan penulis.

13) Rivaldy Putrasena, Reza Akram, Satrio Adhi yang selalu mendukung penulis untuk giat dalam mengerjakan tugas akhir.

14) Teman-teman B/8 yang tidak dapat disampaikan secara satu-persatu yang terus memberi dukungan untuk penulis.

15) Pasta Kangen, Koridor, Koleng, warkop Unair, Kopi Ganes, Blackbarn yang telah menyediakan tempat nyaman dan kondusif bagi penulis dalam mengerjakan tugas akhir.

16) Serta semua pihak yang telah membantu dalam pengerjaan Tugas Akhir ini yang belum mampu penulis sebutkan diatas.

Terima kasih atas segala bantuan, dukungan, serta doa yang telah diberikan. Penulis menyadari bahwa tugas akhir ini masih belum sempurna dan memiliki banyak kekurangan di dalamnya. Oleh karena itu, penulis juga memohon maaf atas segala kesalahan penulis buat dalam buku tugas akhir ini. Penulis membuka pintu selebar-lebarnya bagi pihak yang ingin memberikan kritik maupun saran, serta penelitian selanjutnya yang ingin menyempurnakan karya dari tugas akhir ini. Semoga buku tugas akhir ini bermanfaat bagi seluruh pembaca.

Surabaya, Januari 2018

Penulis

Halaman ini sengaja dikosongkan

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

Halaman ini sengaja dikosongkan

BAB I
PENDAHULUAN

Dalam bab pendahuluan ini akan menjelaskan mengenai latar belakang masalah, perumusan masalah, batasan masalah, tujuan tugas akhir, dan manfaat dari kegiatan tugas akhir. Berdasarkan uraian pada bab ini diharapkan mampu memberi gambaran umum permasalahan dan pemecahan masalah pada tugas akhir.

1.1. Latar Belakang Masalah

Meningkatnya pengguna internet mendorong kepada keberagaman pelanggan pada suatu website e-commerce, terlebih lagi perilaku pelanggan telah berubah begitu drastis dalam beberapa tahun terakhir, pelanggan modern lebih mengerti akan kebutuhannya dan memiliki ekspektasi yang lebih tinggi akan apa yang ditawarkan oleh perusahaan [1][2]. Agar e-commerce memiliki keuntungan yang maksimal dari pelanggannya maka memiliki pengetahuan yang cukup akan pelanggannya merupakan aspek penting agar dapat ditentukan strategi hubungan pelanggan yang tepat. Namun seperti yang disebutkan sebelumnya, saat ini pelanggan memiliki profil yang semakin beragam, dan setiap profil memiliki kebutuhan informasi dan cara penanganan yang berbeda pula [3].

Sebagai upaya untuk memiliki hubungan pelanggan yang baik, maka preferensi dan tingkah laku dari pelanggan harus didapatkan dan dicatat sehingga kedepannya dapat dikembangkan layanan terpersonalisasi yang berfokus pada pelanggan. Saat ini kondisi ketersediaan informasi di internet mengenai pelanggan tidak memiliki struktur yang jelas, oleh karenanya e-commerce memiliki kesulitan dalam membaca dan mencari informasi yang tepat dalam melakukan identifikasi pelanggannya. Personalisasi dan kustomisasi telah menjadi pilihan model bisnis yang terbukti bagi e-commerce, telah banyak e-commerce yang sukses dengan menerapkan layanan

personalisasi dan kustomisasi dengan tujuan menciptakan hubungan yang kuat dengan pelanggannya [4].

Namun sebelum mengembangkan layanan terpersonalisasi pertama-tama e-commerce butuh melakukan pemetaan dan identifikasi terhadap pelanggannya yang diidentifikasikan berdasarkan keinginan, daya beli, perilaku pembelian, kebiasaan pembelian yang serupa, aspek geografis, aspek psikografis, dan aspek yang akan digunakan pada penelitian ini yaitu aspek demografis [1][5].

Penting bagi e-commerce untuk mengetahui siapa pelanggannya, namun kondisi saat ini upaya identifikasi pelanggan masih dilakukan secara manual yaitu dengan menyocokkan setiap pelanggan dengan taksonomi demografi yang telah ditentukan secara satu persatu. Tentu kondisi ini memakan beban kerja dan waktu yang lebih, dan apabila tidak dilakukan maka akan mempengaruhi hubungan pelanggan e-commerce sebab sulit ditentukan strategi hubungan pelanggan karena tidak tersedianya data demografi pelanggan dan tidak dapat mengembangkan layanan personalisasi. Dengan demikian dibutuhkan suatu sistem untuk melakukan identifikasi pelanggan secara otomatis. Permasalahan ini dapat diatasi dengan dilakukan ekstraksi informasi profil pelanggan dengan menerapkan template filling.

Oleh karena itu pada tugas akhir ini fokus pada pengembangan sebuah prototipe yang menjalankan tugas ekstraksi informasi dan identifikasi profil pelanggan berdasarkan demografinya. Ekstraksi informasi pelanggan menggunakan sumber data hasil scraping pada website penyedia layanan konten digital Google Play Store dan menggunakan atribut dari taksonomi demografi yang telah dilakukan pada penelitian sebelumnya.

Pada tugas akhir ini dibangun perangkat lunak yang menerapkan metode template filling, yaitu sebuah pendekatan untuk melakukan ekstraksi dan membuat struktur informasi yang berasal dari teks. Peran dari pendekatan ini adalah untuk menggambungkan informasi dari bermacam-macam informasi yang ada pada sumber data dan mengidentifikasi teks yang

terkait dengan kamus ekstraksi dan membuat informasi menjadi terstruktur [6]. Dengan implementasi ekstraksi informasi menggunakan template filling maka e-commerce tidak perlu melakukan identifikasi pelanggan berdasarkan demografinya secara manual sehingga meringankan beban dalam melakukan tugas identifikasi. Data yang didapat melalui hasil scraping diolah dengan pendekatan rule-based system dan dibuat menjadi terstruktur dengan template filling menghasilkan daftar pelanggan yang telah diidentifikasikan berdasarkan demografinya.

Harapannya dengan dilakukan penelitian ini dapat dijadikan acuan dalam mengembangkan sistem ekstraksi informasi profil demografi pelanggan yang memiliki manfaat dalam meningkatkan efisiensi dalam melakukan identifikasi profil demografi pengguna dan memungkinkan adanya pengembangan layanan terpersonalisasi.

1.2. Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan pada bagian sebelumnya, maka rumusan masalah yang akan diselesaikan pada Tugas Akhir ini adalah sebagai berikut:

1. Bagaimana model template yang dirancang dengan berdasarkan pemetaan ketersediaan data pada e-commerce dan template dari penelitian terdahulu?
2. Bagaimana metode identifikasi menjalankan tugas ekstraksi profil demografi pelanggan?
3. Bagaimana hasil identifikasi profil demografi pelanggan dengan menggunakan template yang dirancang pada penelitian?

1.3. Batasan Masalah

Sesuai dengan deskripsi permasalahan yang telah dijelaskan diatas, adapun batasan permasalahan dari penyelesaian tugas akhir ini adalah sebagai berikut:

a. Ekstraksi informasi profil pengguna berdasarkan demografi.
b. Tugas akhir menggunakan metadata atribut demografi pelanggan yang sudah ada dari penelitian sebelumnya.
c. Prototipe sistem yang dibangun terbatas pada proses ekstraksi informasi demografi user dan template filling.
d. Tugas akhir menggunakan data yang bersumber dari layanan konten digital Google Play Store dengan menggunakan API Google Play Store.

1.4. Tujuan Penelitian

Tujuan dari dilakukannya penelitian ini adalah untuk membuat sebuah prototipe perangkat lunak yang melakukan ekstraksi informasi pelanggan dan secara otomatis mengidentifikasi profil demografi pelanggan, menghasilkan data pelanggan yang telah diidentifikasikan berdasarkan profil demografinya.

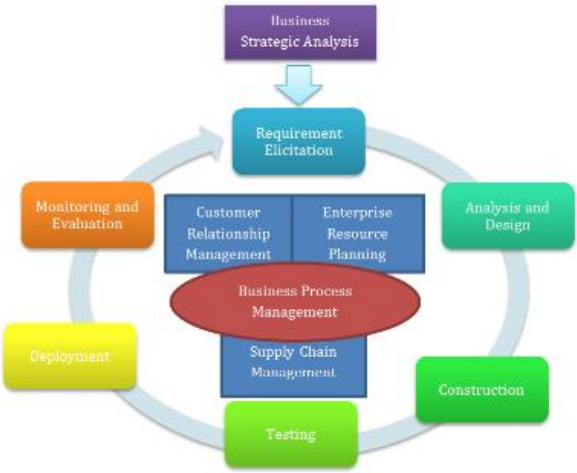
1.5. Manfaat Penelitian

Adapun manfaat yang dapat diperoleh yang dibedakan menjadi dua belah sudut pandang sebagai berikut:

1. Bagi pemilik e-commerce, memberikan tambahan fitur yang dapat meringankan beban e-commerce dalam mengidentifikasi pelanggan berdasarkan demografinya.
2. Bagi developer, meningkatkan eksposur aplikasi kepada konsumen yang sesuai dengan profil pelanggan aplikasi tersebut.

1.6. Relevansi

Laboratorium Sistem Enterprise (SE) Jurusan Sistem Informasi ITS memliki empat topik utama yaitu customer relationship management (CRM), enterprise resource planning (ERP), supply chain management (SCM) dan business process management (BPM) seperti yang terdapat pada Gambar 1.1 Dalam tugas akhir yang dikerjakan oleh penulis mengambil customer relationship management (CRM) sebagai topik utama. Mata kuliah yang berkaitan dengan CRM adalah Manajemen Rantai Pasok dan Hubungan Pelanggan (MRPHP).



Gambar 1.1. Kerangka kerja riset laboratorium sistem enterprise

Halaman ini sengaja dikosongkan

BAB II
TINJAUAN PUSTAKA

Bagian ini akan memberikan penjelasan mengenai penelitian maupun studi literatur sebelumnya yang berkaitan dan dijadikan sebagai acuan selama pengerjaan tugas akhir, serta landasan teori yang berkaitan dengan tugas akhir yang dapat membantu pemahaman selama pengerjaan tugas akhir ini.

2.1. Penelitian Sebelumnya

Selama pengerjaan tugas akhir ini, terdapat beberapa penelitian yang telah dilakukan sebelumnya yang dapat dijadikan sebagai bahan kajian maupun referensi untuk studi literatur. Penelitian tersebut lalu dikaji untuk dilihat dari gambaran umum, tujuan hasil, dan keterkaitannya dengan penelitian tugas akhir ini. Hasil dari kajian tersebut dapat kita lihat pada tabel 2.1 berikut.

Tabel 2.1 Penelitian Sebelumnya

Publikasi	Knowledge and Systems Engineering, 2012
Judul Paper	A Combined Approach for Disease/Disorder Template Filling
Penulis	Nghia Huynh, Quoc Ho.
Gambaran umum penelitian	Disease/Disorder template filling merupakan pekerjaan ekstraksi relasi antar variabel yang rumit. Untuk melakukan pekerjaan ini dibutuhkan lebih dari satu metode untuk menyelesaikannya. Tujuan

	dari penelitian ini adalah membuat sebuah usulan untuk melakukan disorder template filling dengan menggabungkan 3 metode yaitu rule-based, regular expression (Regex), dan machine learning [7].
Keterkaitan penelitian	Pada penelitian ini terdapat template terdiri dari 10 atribut yang berbeda. Penelitian ini membuat sebuah pendekatan untuk melakukan pengisian template disease/disorder dengan mengacu 10 atribut secara otomatis dan membangun sistem dengan mencakup 3 pendekatan dalam memprediksikan nilai dari template. Masing-masing atribut memiliki penggunaan metode yang mencakup metode rule-based, machine-learning, dan regular expression.

Publikasi	Electrical Engineering and Informatics, 2011

Judul Paper	Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application.
Penulis	Sri Krisna Endarnoto, Sonny Pradipta, Anto Satriyo Nugroho, James Purnama.
Gambaran umum penelitian	Pada penelitian ini dikembangkan sistem yang menyediakan berita lalu lintas dari sumber terpercaya dengan menerapkan ekstraksi informasi dan template filling. Metodologi pada penelitian ini diawali mengambil tweet dari media sosial Tweeter dan menyimpannya ke sebuah tabel database. Selanjutnya dilakukan tokenization terhadap tweet, yaitu memisahkan kata-kata yang terdapat menjadi token yang terpisah-pisah. Selanjutnya dilakukan Part Of Speech Tagging (POS) yang memberi label terhadap setiap token yang dihasilkan di tahapan

	sebelumnya dengan mengacu nama label POS yang telah diatur. Dengan aturan yang telah didefinisikan pada database, proses tagging dilakukan dengan mencari token yang sesuai dengan salah satu koleksi kosakata pada database dan menyesuaikannya dengan nama POS. Bersama dengan tahap POS, dilakukan sentence analysis, yaitu dari setiap token yang telah diberi label dilakukan analisa oleh sistem dengan menggunakan aturan yang telah didefinisikan, pada penelitian ini aturan digunakan untuk mendapatkan informasi mengenai waktu, tempat asal, tempat tujuan, dan kondisi lalu lintas. Keempat informasi tersebut disimpan ke database, atribut-atribut dari informasi tersebut pada dasarnya adalah template dan pada tahap ini (template filling) database akan

	diisikan dengan informasi yang telah diekstraksi [8].
Keterkaitan penelitian	Penelitian ini memberikan sudut pandang yang menyajikan metodologi dalam melakukan ekstraksi informasi, tokenization, POS, dan template filling. Metodologi pada penelitian ini dapat diterapkan pada pengerjaan tugas akhir.

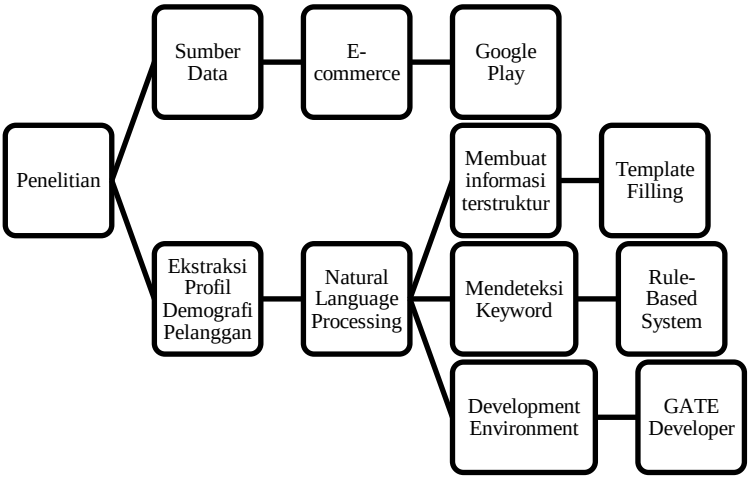
Publikasi	Mathematic and Statistic, 2008
Judul Paper	Ontology-Based User Modeling for E-Commerce System
Penulis	Weilong Liu, Fang Jin, Xin Zhang
Gambaran umum penelitian	Penelitian ini menjelaskan pendekatan yang dilakukan dalam melakukan pemodelan *user*, yaitu dengan menerapkan *ontology based user modelling*. *User modelling* pada umumnya menjelaskan profil dari pengguna dan merupakan pendekatan untuk mendapatkan *preference* dari pengguna. Pada penelitian ini disampaikan bahwa

	user modelling merupakan kegiatan kunci dalam operasi marketing dan memungkinkan adanya layanan personalisasi pada suatu *e-commerce*. *User modelling* diperlukan untuk membentuk pengetahuan mengenai pelanggan *e-commerce.* Dalam melakukan *user modelling*, pada penelitian ini juga melibatkan pemetaan demografi dengan mengacu *form register* yang spesifikasinya didesain oleh Information Management System Learner Information Package (IMS LIP). Adapun metadata profil pengguna sebagai berikut: [20] • Identity: Menjelaskan identitas pengguna sistem. • Gender: Menjelaskan jenis kelamin dari pengguna. • Age: Menjelaskan umur dari pengguna.

	• Profession: Menjelaskan profesi dari pengguna. • Email: Menjelaskan bentuk kontak dari pengguna. • Phone: Menjelaskan bentuk kontak dari pengguna. • Address: Menjelaskan bentuk kontak dari pengguna. • Product_preference: Daftar *preference* produk dari pengguna. • User_rank: Evaluasi ranking pengguna berdasarkan seberapa sering menggunakan sistem.
Keterkaitan penelitian	Penelitian ini memiliki relevansi yang terkait dengan tugas akhir yaitu mengenai pemetaan pelanggan e-commerce. Penelitian ini menjabarkan ontologi pengguna e-commerce yang akan digunakan sebagai atribut template pada tugas akhir ini.

2.2. Dasar Teori

Bagian ini akan membahas teori dan konsep yang berkaitan dengan penelitian tugas akhir. Relevansi dari dasar teori yang dibahas pada penelitian ini dapat dilihat pada bagan berikut ini:



Dasar teori disusun untuk menjawab pertanyaan berikut:

1. Dari manakah sumber data didapat?
2. Sudut pandang apakah dalam melakukan segmentasi pelanggan?
3. Teknologi apakah yang diterapkan agar komputer mampu mengerti dan memahami makna bahasa manusia dan memberikan respon yang sesuai?
4. Bagaimana cara merubah teks tidak terstruktur menjadi dalam bentuk yang terstruktur?

5. Metode apakah yang tepat dalam melakukan deteksi keyword?
6. Bagaimana menguji hasil pendeteksian sistem?

2.2.1. *E-Commerce*

Istilah e-commerce merujuk kepada pembelian, penualan barang serta jasa, pelayanan pelanggan, kolaborasi dengan mitra bisnis, e-learning, dan transaksi perusahaan yang melibatkan peran elektronik apapun. Terdapat berbagai bentuk e-commerce tergantung pada tingkat digitalisasi (tranformasi dari pekerjaan fisik ke digital) yang dilibatkan. Tingkat digitalisasi dapat berhubungan dengan (1) produk dan/atau jasa yang dijual, (2) proses, dan (3) pelaku pengirimannya. Dalam menjelaskan kemungkinan dari 3 dimensi tersebut sudah terdapat kerangka yang dibuat oleh Choi, et al. Perusahaan yang murni fisik disebut "brick and mortar" yang artinya perusahaan melakukan operasinya sepenuhnya secara tradisional. Jika terdapat paling tidak satu dimensi digital maka kondisi tersebut akan dianggap sebagai e-commerce parsial. Dan jika ketiga dimensi tersebut dilakukan secara digital maka kondisi tersebut dianggap sebagai e-commerce murni [9].

Transaksi e-commerce dapat dilakukan antara berbagai pihak. Adapun jenis e-commerce mencakup business-to-business (B2B), collaborative business, business-to -consumers (B2C), consumer-to-consumer (C2C), consumer-to-business (C2B), intraorganisational, goverment-to-citizen (G2C) [9].

Pada penelitian ini digunakan Google Play sebagai studi kasus yang termasuk jenis business-to-consumers (B2C) dimana dalam transaksinya penjual berupa perusahaan dan pembeli adalah perorangan. Google Play juga merupakan bentuk e-commerce yang berbentuk e-commerce dimana produk, proses, dan transaksinya dilakukan secara digital.

2.2.2. Google Play

Google Play yang pada awalnya bernama Android Market adalah layanan konten digital untuk membeli atau

mengunduh aplikasi Android secara gratis. Saat ini sudah terdapat lebih dari jutaan aplikasi terdapat pada layanan Google Play yang terdiri dari aplikasi berbayar dan gratis. Adapun aplikasi yang terdapat di Google Play mencakup aplikasi software, musik, film, dan buku [10].

Google Play diperkenalkan pertama kali pada tanggal 28 Agustus 2008 dan dapat diakses oleh pengguna pada tanggal 22 Oktober 2008. Layanan ini mulai dapat mendukung aplikasi berbayar pada anggal 13 Febuari 2009 hanya untuk Amerika Serikat dan Inggris yang selanjutnya berekspansi ke 29 negara pada tanggal 30 September 2010 [11].

Pengguna dari aplikasi Android terdiri dari berbagai latar belakang mulai dari pelajar, pekerja, hingga masyarakat senior, sehingga mendorong pembuatan aplikasi yang mempertimbangkan penggunaan oleh berbagai latar belakang pengguna yang luas. Dengan peluncurannya ponsel Android one, semakin banyak masyarakat yang memiliki akses untuk memiliki ponsel dengan platform Android yang berarti akses untuk mengunduh aplikasi berbasis Android semakin besar dan menumbuhkan komunitas pengembang aplikasi [10].

Pada penelitian ini digunakan Google Play Store sebagai studi kasus. Penelitian ini akan menggunakan fitur ulasan pengguna sebagai sumber data untuk dilakukan analisis demografi. Pada fitur ulasan pengguna terdapat berbagai informasi seperti nama pemberi ulasan, tanggal ulasan, dan komentar. Contoh fitur ulasan pengguna dapat dilihat sebagai berikut:

Gambar 2.1 Contoh Ulasan Pengguna

Dengan mengacu informasi yang terdapat pada fitur ulasan pengguna seperti gambar 2.1 diatas, maka dapat dilihat dicantumkan informasi nama yang merupakan salah satu atribut yang dapat menjelaskan demografi pelanggan aplikasi.

2.2.3. *Demographics*

Demographics atau demografi adalah studi matematik dan statistik terhadap komposisi dan distribusi populasi manusia. Pada umumnya studi ini menganalisa dinamika perubahan yang terjadi pada populasi manusia yang mencakup lima proses yaitu fertilitas (kelahiran), mortalitas (kematian), perkawinan, migrasi dan mobilitas sosial. Dengan kata lain demografi meliputi studi mengenai ukuran, struktur, distribusi dari populasi serta bagaimana jumlah penduduk berubah setiap waktu akibat kelahiran, migrasi, serta penuaan. Dalam studi demografi adapun istilah demographic analysis yaitu merupakan analisa yang merujuk masyarakat secara keseluruhan ataupun kelompok tertentu yang didasarkan kriteria seperti pendidikan, kewarganegaraan, agama, atau etnisitas tertentu [12]. Adapun atribut demografi umum mencakup kepadatan penduduk, etnis, tingkat edukasi, tingkat

kesehatan, status ekonomi, agama, dan banyak aspek-aspek lainnya. Pada penelitian ini digunakan kriteria demografi dalam mengidentifikasi pelanggan dan digunakan dalam menentukan atribut-atribut template.

2.2.4. *Natural Language Processing*

Natural Language Processing (NLP) atau pengolahan bahasa alami adalah bidang studi dan terapan mengenai bagaimana komputer dapat digunakan untuk mengerti dan memanipulasi teks natural language dan menghasilkan output yang diinginkan. NLP dikembangkan untuk menambah pengetahuan mengenai bagaimana manusia memahami dan menggunakan bahasanya sehingga dapat dikembangkan teknik dan tools yang tepat dalam membuat sistem komputer yang dapat memahami dan memanipulasi natural language dan melakukan tugas yang diinginkan [13].

Algoritma NLP pada umumnya berdasarkan algoritma machine learning. Dalam pembuatan aturan-aturan pengolahan, NLP dapat digunakan dengan bergantung pada machine learning dalam mempelajari aturan-aturan secara otomatis yang dilakukan dengan menganalisa sekumpulan contoh dan menghasilkan kesimpulan statistik. Semakin banyak data yang dianalisa maka model yang dihasilkan menjadi lebih akurat. Pada penelitian ini, pendekatan yang digunakan adalah pendekatan dengan menggunakan finite-state cascade. Pendekatan ini akan mengambil masukan berupa features yang didapatkan dari data tertentu.

2.2.5. *Information Extraction*

Information Extraction atau Ekstraksi Informasi adalah teknologi yang bertugas melakukan ekstraksi informasi pada dokumen teks yang tidak terstruktur dan/atau semi-terstruktur dan mengubahnya menjadi informasi dalam bentuk terstruktur. Teknologi ini termasuk bidang ilmu untuk pengolahan bahasa alami (natural language processing). Tujuan umum dari

ekstraksi informasi adalah adanya komputasi yang dilakukan pada teks yang tidak terstruktur. Tujuan lebih spesifiknya adalah untuk memungkinkan adanya penalaran logis dalam membuat kesimpulan dari konten input data [14].

Cara kerja dari teknologi ini adalah mencari informasi yang relevan dengan domain yang telah ditentukan dan tidak mempedulikan informasi yang tidak relevan. Secara umum, proses ekstraksi informasi terdiri dari dua tahap, yaitu mengidentifikasi data yang relevan kemudian menyimpan data tersebut kedalam bentuk terstruktur untuk digunakan di kemudian.

Proses ekstraksi informasi secara manual dapat dilakukan dengan membaca teks dari sebuah dokumen teks, mengidentifikasi data apa saja yang relevan, kemudian data yang relevan disimpan kedalam basis data. Proses ini membutuhkan usaha dan waktu yang besar karena berhubungan dengan informasi tekstrual yang berskala besar. Solusinya adalah dengan mendefinisikan aturan ekstraksi yang digunakan untuk mengekstrak informasi yang diinginkan dari sebuah dokumen teks sehingga adanya otomasi dalam menjalankan proses ekstraksi informasi. Pada penelitian ini dilakukan ekstraksi informasi pada teks yang berkaitan dengan atribut demografi. Keyword atribut profil demografi yang terdapat pada teks yang tidak terstruktur diidentifikasi dan diubah menjadi terstruktur.

2.2.6. *Template Filling*

Template Filling adalah sebuah pendekatan efisien yang dilakukan untuk melakukan ekstraksi struktur informasi kompleks yang bersumber dari text. Pendekatan ini memiliki peran penting pada sistem information extraction (IE) untuk melakukan penggabungan informasi dari beberapa kalimat yang ada untuk mengidentifikasi peran isi dari sumber informasi yang bentuknya bisa berupa dokumen dan/atau teks. Tugas dari template filling dapat dipahami dari namanya, yaitu template dan fill. Sebuah template, atau juga sebuah skema

abstrak dapat didefinisikan sebagai penentu pemilihan dan mempersempit domain of interest menjadi informasi umum dari interest dan bentuk output baru informasi. Fill rules mendefinisikan template dengan menentukan aturan-aturan dalam melakukan ekstraksi informasi umum untuk mengisi template dan secara formal dijalankan sebagai pedoman ekstraksi informasi [6].

Desain struktur template untuk penerapan IE tergantung dari domain of interest dan lingkungan dari tugas. Dalam template filling, data menjelaskan informasi dari interest. Pada tabel 2.3 berikut ini adalah sebuah ilustrasi penggunaan template filling:

Text	Template
Sentence 1 Nama saya adalah Tony, saya berumur 25 tahun dan tinggal di kota Jakarta. Saat ini saya merupakan berprofesi sebagai karyawan swasta.	Sentence 1 Nama: Tony Umur: 25 tahun Tempat Tinggal: Jakarta Profesi: karyawan swasta

Gambar 2.2 Contoh Penerapan Template Filling

Pada penelitian ini digunakan template filling dalam menyimpan informasi secara terstruktur sesuai dengan aturan atribut yang ditentukan. Pendekatan template filling yang digunakan adalah finite-state cascade.

2.2.7. *Rule-Based System*

Rule-based System atau sistem berbasis aturan merupakan metode untuk menyimpan atau memanipulasi

pengetahuan untuk mengintepretasikan informasi yang berguna bagi penggunanya. Umumnya sistem ini diimplementasikan dengan sistem Artificial Intelligence. Suatu aturan terdiri dari bagian yaitu:

1. Antacedent, yaitu bagian yang mengekspresikan situasi atau premis. Pernyataan berawalan IF.
2. Consequent, yaitu bagian yang menyatakan suatu tindakan tertentu atau konklusi yang diterapkan jika situasi atau premis bernilai benar. Pernyataan berawalan THEN.

Pada umumnya sebuah aturan dapat memiliki gabungan beberapa antecedent dengan kata kunci AND (konjungsi), OR (disjungsi) ataupun kombinasi keduanya [14]. Rule-based system mencakup satuan aturan yang menjelaskan pengetahuan dari domain terkait. Bentuk umum dari aturan dapat dilihat pada tabel 2.4 berikut:

If cond1

And cond2

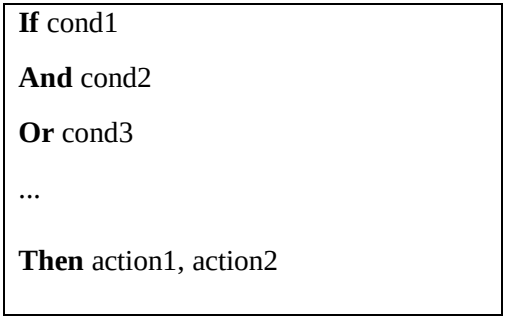
And cond3

...

Then action1, action2

Gambar 2.3 Contoh Rule-based System 1

Bentuk kondisi yang diatas adalah antecedents dilakukan evaluasi berdasakan apa yang diketahui untuk menyelesaikan permasalahan. Beberapa sistem terkadang juga menggunakan disjungsi dalam antecedents,contohnya sebagai berikut [15]:



Gambar 2.4 Contoh Rule-based System 2

Penggunaan rule-based system dapat memberikan kelebihan seperti berikut [15].

1) Homogenitas

Karena memiliki sintaks yang seragam, makna dan interpretasi dari masing-masing aturan dapat dengan mudah dianalisis.

2) Kesederhanaan

Karena sintaks sederhana, mudah untuk memahami makna dari aturan. Ahli domain seringkali dapat memahami aturan tanpa penerjemahan yang eksplisit. Aturan sehingga dapat mendokumentasikan diri sampai batas yang baik.

3) Independensi

Ketika menambahkan pengetahuan yang baru tidak perlu khawatir tentang dimana aturan itu akan ditambahkan, atau apakah ada interaksi dengan aturan lainnya. Secara teori, setiap aturan adalah bagian independen dari pengetahuan tentang domain tersebut. Namun, dalam prakteknya, hal ini tidak sepenuhnya benar.

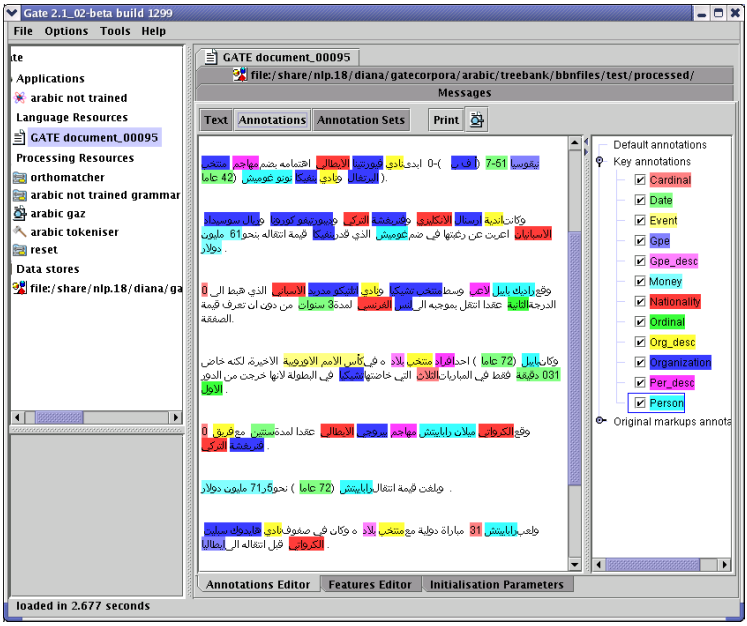
4) Modularitas

Independensi aturan mengarah ke modularitas dalam rule base. Prototipe sistem dapat diciptakan cukup cepat dengan

membuat beberapa aturan. Hal ini dapat ditingkatkan dengan memodifikasi aturan berdasarkan kinerja dan menambahkan aturan baru.

Pada penelitian ini digunakan rule-based system untuk melakukan intepretasi dari data yang diterima. Yaitu input yang diterima diubah menjadi output demografi pelanggan yang terstandar.

2.2.8. GATE Developer



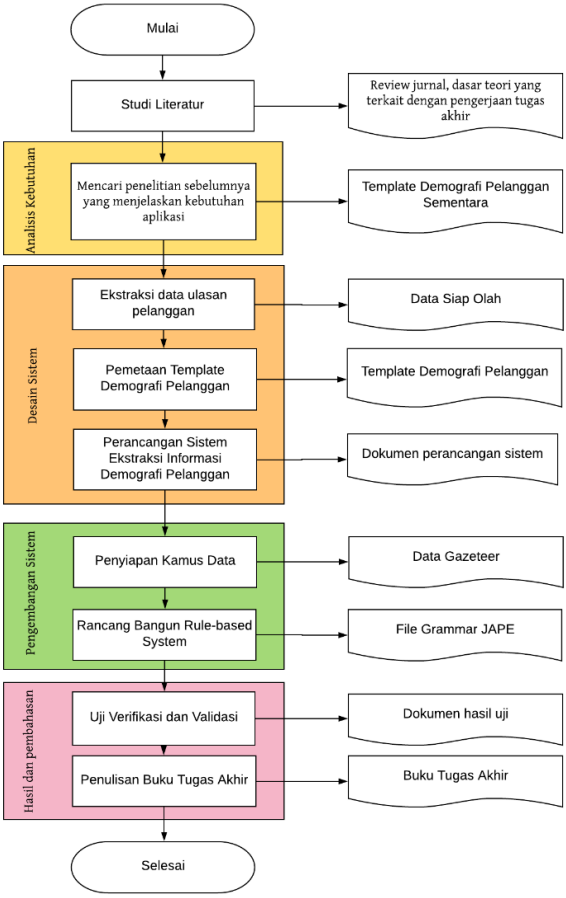
Gambar 2.5 GUI GATE Developer

GATE Developer adalah lingkungan pengembangan yang menyediakan seperangkat alat interaktif grafis yang kaya untuk pembuatan, pengukuran, dan pemeliharaan komponen perangkat lunak untuk memproses bahasa manusia. GATE Developer adalah perangkat lunak open source, tersedia di bawah GNU Lesser General Public License 3.0.

Penggunaan lingkungan pengembangan ini menggunakan struktur data dan algoritma khusus seperti grafik anotasi, *finite state machines*, atau *support vector machines*. GATE Developer membantuk pembuatan struktur kompleks tersebut, menyediakan visualisasi dari pengolahan hasil, dan menyediakan pengukuran ketepatannya relatif terhadap hasil manual atau semi-otomatis.

BAB III
METODOLOGI PENELITIAN

Pada bab ini menjelaskan terkait metodologi yang akan digunakan sebagai panduan untuk menyelesaikan penelitian tugas akhir ini.



Gambar 3.1 Bagan Metodologi

3.1. **Tahapan**

Penjelasan setiap tahapan dari metodologi adalah sebagai berikut:

3.1.1. Studi Literatur

Pada tahap ini penulis melakukan identifikasi masalah terkait dengan upaya e-commerce dalam memetakan pelanggannya. Masalah yang ditemukan akan ditetapkan menjadi latar belakang permasalahan, rumusan masalah, batasan masalah, tujuan dan manfaat dari tugas akhir. Studi literatur dilakukan dengan cara membaca refrensi buku dan penelitian yang telah dilakukan sebelumnya. Tujuan utama dari dilakukannya tahapan ini agar penulis dapat memahami dasar teori yang berhubungan dengan permasalahan agar dapat mempermudah dalam mengembangkan solusi yang tepat. Pada tahapan ini juga dilakukan kajian Application Programmable Interface (API) text mining yang sudah beredar di internet untuk melihat bagaimana cara kerja metode yang diterapkan, dan juga dilakukan pemetaan data-data yang tersedia pada website Google.

3.1.2. Analisis Kebutuhan

Pada tahapan ini dilakukan analisis kebutuhan untuk menjelaskan kemampuan aplikasi seperti apa yang ingin dicapai. Pada tahapan ini dilakukan observasi dan pengamatan dalam membangun template. Didapatkan kajian kebutuhan dari penelitian yang telah dilakukan sebelumnya berjudul "Ontology-Based User Modeling for E-Commerce System" [20]. Pada penelitian ini dijelaskan metadata apa saja yang diperlukan dalam memetakan profil demografi pelanggan e-commerce. Metadata yang dicakup pada penelitian ini merupakan hasil desain dengan menyesuaikan spesifikasi Information Management System Learner Information Package (IMS LIP), yaitu sebuah standar yang mempermudah interoperabilitas sistem informasi berbasis internet dalam

mendukung pembelajaran di internet. Selain itu pada tahapan ini dilakukan identifikasi peluang apa saja yang terdapat dalam melakukan pemetaan profil pelanggan dengan melihat ketersediaan data dan mencocokannya metode yang dapat diterapkan dalam mengidentifikasi.

3.1.3. Desain Sistem

Pada tahapan ini dilakukan desain pembuatan sistem lunak yang mecakup struktur data, arsitektur perangkat lunak, dan prosedur pengodean. Di tahapan ini ditetapkan API yang akan digunakan dalam membangun aplikasi. Untuk pengambilan data dengan menggunakan Google Play API yang telah dibuat oleh Natanael Yabes Wirawan [17]. Pada penelitian ini menggunakan atribut template sesuai dengan penelitian yang berjudul "An Ontology-Supported User Modeling Technique with Query Templates for Interface Agents" [4]. Selanjutnya dilakukan pemetaan template demografi pelanggan dengan membandingkan atribut template dari penelitian sebelumnya dengan ketersediaan data yang disediakan Google Play API menghasilkan template demografi pelanggan yang akan dijadikan acuan pengembangan sistem. Berdasarkan template yang dihasilkan dari pemetaan, dilakukan perancangan sistem ekstraksi informasi demografi pelanggan yang mencakup perancangan kamus data dan perancangan *rule-based system.*

3.1.4. Pengembangan Sistem

Pada langkah ini dilakukan penyiapan kamus data yaitu mengumpulkan data-data yang akan dijadikan kamus data sistem dalam melakukan ekstraksi informasi demografi pelanggan. Kamus data dikumpulkan dari berbagai sumber tergantung dari atribut terkait. Langkah ini menghasilkan data *gazeteer*, yaitu kamus data yang akan dijadikan acuan sistem dalam mengidentifikasi demografi pelanggan. Pada tahapan pengembangan sistem juga dilakukan rancang bangun *rule-based system* yaitu logika sistem dalam mengolah input data

pelanggan bagaimana mengintepretasikan informasi demografinya. *Rule-based system* pada penelitian ini dibangun menggunakan lingkungan pengembangan GATE Developer, dokumen *rule* akan disimpan dalam format JAPE yang khusus digunakan untuk GATE Developer.

3.1.5. Hasil dan Pembahasan

Pada tahapan ini akan dilakukan dua jenis pengujian, diawali dengan dilakukannya pengujian verifikasi untuk melihat apakah metode penelitian dapat menjalankan tugasnya dalam mendeteksi pola-pola atribut demografi pelanggan beserta tingkat performa yang dihasilkan oleh masing-masing rule. Verifikasi dilakukan dengan memasukkan muatan data dengan masing-masing data yang telah diberi label kemudian diproses dengan sistem. Selanjutnya dibandingkan hasil anotasi dihasilkan oleh sistem yang bernilai benar sesuai dengan label muatan input data.

Selanjutnya dilakukan pengujian validasi yaitu melihat apakah aplikasi dapat memberi bantuan dalam melakukan pemetaan demografi pelanggan secara otomatis sesuai dengan tujuan penelitian. Setelah semua proses dalam penelitian telah berjalan dengan baik dan teruji selanjutnya dilakukan proses penyusunan laporan tugas akhir. Validasi dilakukan dengan menggunakan muatan input data dari Google Play dan berdasarkan muatan data tersebut dilakukan pengamatan hasil anotasi yang dihasilkan oleh sistem. Dari hasil anotasi yang dihasilkan dilakukan rekap data untuk melihat informasi demografi pelanggan.

BAB IV
PERANCANGAN

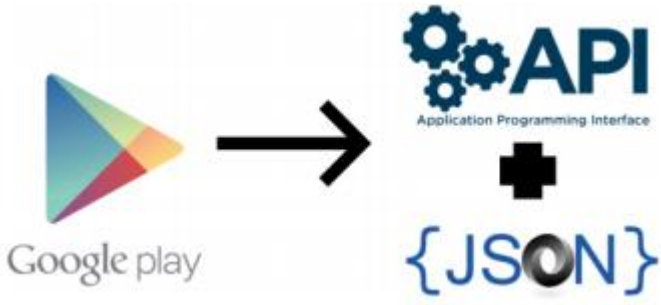
Untuk dapat memberikan gambaran apa-apa saja yang dikerjakan pada implementasi penelitian tugas akhir, pada bab ini menjelaskan perancangan penelitian tugas akhir yang meliputi subyek dan obyek penelitian, pemilihan subyek dan obyek penelitian serta bagaimana penelitian akan dilakukan.

4.1. Perancangan pengambilan data

Pada bagian ini akan dijelaskan mengenai perancangan pengambilan data yang bertujuan sebagai panduan ketika akan melakukan implementasi pada penelitian ini.

4.1.1. Perancangan Pengambilan Data

Pengambilan data dilakukan sebagai sumber data yang akan diolah dengan metode template filling sehingga dapat menghasilkan data yang lebih terstruktur. Untuk melakukan proses pengambilan data digunakan API yang telah dirancang sesuai dengan struktur situs Google Play sehingga bentuk data yang diambil sudah lebih terstruktur dan memudahkan pengerjaan pada penelitian ini. Adapun API yang digunakan dalam proses pengambilan data yang dapat diakses dengan alamat web gooplayapi.herokuapp.com, API ini dibangun dalam penelitian tugas akhir jurusan Sistem Informasi oleh Natanael Yabes W. Data yang diambil merupakan data *reviews* yaitu mencakup ulasan pengguna aplikasi pada Google Play. Berikut merupakan alur pengambilan data menggunakan Google Play API.



Gambar 4.1 Alur pengambilan data Google Play API

Adapun data yang dibutuhkan pada penelitian berupa data-data ulasan produk aplikasi. Pengambilan data pada Google Play API dilakukan secara *real time* melalui proses scraping berbasis pemrograman PHP dengan struktur tabel seperti tabel 4.1 berikut:

Tabel 4.1 Deskripsi data atribut tabel reviews

Nama Tabel	Nama Kolom	Tipe Data	Keterangan
reviews	Id	Int(11)	Primary Key untuk tabel reviews
reviews	App_id	Int(11)	Foreign key untuk menghubungkan ke tabel Applications
reviews	Username	Varchar(45)	Nama Username dari pereview aplikasi
reviews	Userimage	Varchar(100)	Gambar Profile Picture dari pereview aplikasi

Nama Tabel	Nama Kolom	Tipe Data	Keterangan
reviews	Date	Varchar(50)	Tanggal dari ulasan yang dibuat oleh pereview aplikasi.
reviews	url	Varchar(200)	Alamat URL dari aplikasi yang direview
reviews	score	Int(11)	Nilai atau score yang diberikan oleh pereview aplikasi
reviews	title	Varchar(45)	Judul dari review yang diberikan
reviews	text	text	Komentar dari pereview aplikasi

Teknik yang digunakan untuk pengambilan data dari Google Play API adalah dengan membangun program scraping berbasis pemrograman PHP yang kemudian disimpan kedalam basis data MySQL. Pada penelitian ini hanya digunakan dua tabel yaitu tabel reviews yang dicakup pada butir Tabel 4.1, dan digunakan juga tabel produk sebagai refrensi atau foreign key ID aplikasi. Proses pengambilan data dari API + JSON memiliki alur proses yang dapat dilihat pada Gambar 4.2:

Gambar 4.2 Alur pengambilan data Google Play API

Dari seluruh data yang telah didapatkan dan disimpan akan dijadikan sumber data dalam membangun *rules* dalam mendefinisikan nilai dari atribut demografi pelanggan. Tabel 4.2 berikut menunjukkan daftar data yang disediakan pada Google Play API:

Tabel 4.2 Data review yang disediakan Google Play API

Nama Atribut	Keterangan
ID	Merupakan ID dari user Google Play Store
userName	Teks dari user name pereview aplikasi Google Play
userImage	Gambar *profile picture* dari pereview aplikasi
date	Tanggal dari komentar
url	URL dari komentar pereview
Score	Nilai yang diberikan oleh pereview terhadap aplikasi
Title	Judul dari komentar pereview
Text	Isi dari komentar pereview terhadap aplikasi

Sebagai gambaran, berikut ini merupakan sampel data review pelanggan yang disediakan pada Google Play API:


```
{
 id: "gp:AOqpTOGK6kbXYBd0qjyu2sgq_smRjvapPmUSDyPn0LHoKF5NNEpbzkNOu5Puzns0FgsH49r4P96EXGgp1Xes79U",
 userName: "Sabir Khan",
 userImage: "https://lh3.googleusercontent.com/-WTnR5Z5KNjw/AAAAAAAAAAI/AAAAAAAAAAA/ACnBePY80u2HFvXBlcxcrZJ6u6aa59fuOg/w96-h96-p/photo.jpg",
 date: "October 11, 2017",
 url: "https://play.google.com/store/apps/details?id=com.netflix.mediaclient&reviewId=Z3A6QU9xcFRPR0s2a2JYWUJkMHFqeXUyc2dxX3NtUmp2YXBQbVVTRHlQbjBMSG9LRjVOTkVwYnprTk91NVB1em5zMEZnc0g0OXI0UDk2RVhHZ3AxWGVzNzlV",
 score: 5,
 title: "",
 text: "Awesome"
}
```


Gambar 4.3 Sampel Data Review Pelanggan

4.1.1. Perancangan Praproses Data

Data yang telah disimpan pada database selanjutnya akan diunduh dan disimpan dalam format .csv (*comma-separated values*) agar selanjutnya dapat diproses pada aplikasi GATE Developer. Terdapat satu proses sebelum data diunggah dan diproses pada aplikasi GATE Developer yaitu dilakukan persiapan data agar format konten dari file .csv dapat diproses

secara maksimal. Tahapan ini dilakukan bertujuan agar aplikasi GATE dapat mendeteksi *token* secara terpisah-pisah dan menghilangkan karakter-karakter yang bersifat *noise* pada *token.* Adapun karakter noise yang terdapat antara lain tanda kutip (“) dan tanda titik koma (;) seperti yang telah ditunjukkan pada Gambar 4.4:

"2";"1";"Karen Powers";"https://lh5.googleusercontent.com/-pVqcdWLIhOk/AAAAAAAAAAI/AAAAAAAAAAA/AHalGhog4nQaAiXSRFv3u_5Fd5XkV";"May 21, 2017";"https://play.google.com/store/apps/details?id=com.netflix.mediaclient&reviewId=Z3A6QU9xcFRPSFR4VXpBa29Ma1FCX202R0lQMEs0bC1EbEp2bVphenowLXZReUZjZTZyUzdiWkRLV3VuQm5oVF9FanFfbDFCSnU2OXNVa29SMFNGTGJ3R";"5";;"Netflix rocks!"

Gambar 4.4 Data Sebelum Diproses

Apabila kondisi data yang belum diproses seperti diatas langsung diunggah dan diproses pada aplikasi GATE Developer maka aplikasi tidak dapat mendeteksi token karena tidak terdapat spasi pada setiap kata dan terdapat *noise* dari setiap kata seperti karakter kutip (“) dan titik koma (;) yang akan mengganggu proses pendeteksian pada setiap token. Maka dari itu data perlu diproses sehingga menjadi bentuk token yang terpisah-pisah dan bersih dari karakter noise seperti yang akan ditunjukkan pada bab implemetasi butir 5.4.

4.2. Identifikasi Atribut Demografi Pelanggan

Pada bagian ini akan dijelaskan bagaimana dan sumber penetapan atribut apa saja yang diperlukan untuk mengidentifikasi demografi pelanggan pada platform Google Play. Pada penelitian ini identifikasi atribut demografi pelanggan dilakukan dengan mengacu penelitian yang telah dilakukan sebelumnya yaitu dengan mengacu penelitian berjudul *Ontology-Based User Modeling for E-Commerce*

System yang ditulis oleh Weilong Liu, Fang Jin, dan Xin Zhang [4].

Pada penelitian diatas disampaikan bahwa data demografis dibentuk secara eksplisit pada sistem e-commerce yaitu ketika user melakukan pendaftaran pada sistem e-commerce yang mencakup kegiatan pengisian *form* data user. Konten dari *form* pendaftaran user dirancang sesuai dengan spesifikasi *Information Management System Learner Information Package* (IMS LIP). Pada tabel 4.3 menunjukkan daftar atribut demografi user yang perlu didefinisikan oleh e-commerce berdasarkan penelitian diatas.

Tabel 4.3 Atribut Demografi User

Atribut	**Keterangan**
Identity	Merepresentasikan identitas user pada sistem e-commerce
Gender	Jenis kelamin dari user
Age	Umur dari user
Profession	Profesi dari user
Email	Bentuk kontak untuk menghubungi user
Address	Bentuk kontak untuk menghubungi user
Phone	Bentuk kontak untuk menghubungi user
Product_preference	Daftar preferensi produk dari user
User_rank	Nilai user rank berdasarkan frekuensi user dalam mengakses sistem e-commerce

Setelah dilakukan identifikasi atribut demografi pelanggan maka selanjutnya akan dilakukan pemetaan ketersediaan data untuk melihat potensi atribut apa saja yang

dapat diidentifikasi. Tahapan ini akan dibahas pada bab implementasi.

4.3. Perancangan Kamus Data

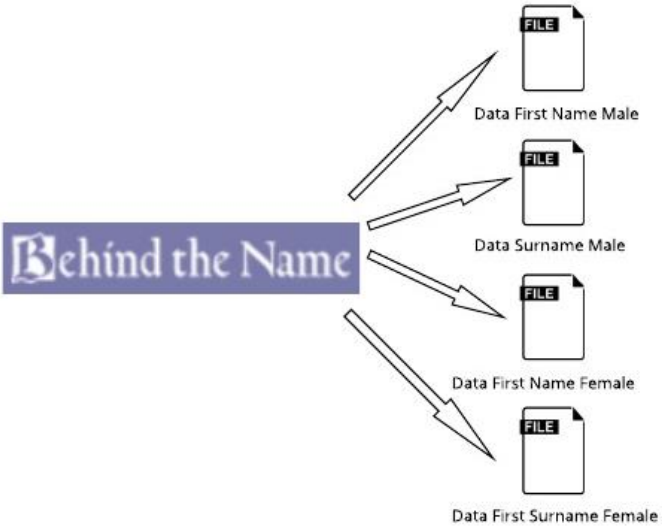
Tahapan perancangan kamus data mencakup hal-hal apa saja yang perlu dilakukan dalam menyediakan data yang akan menjadi acuan sistem dalam mengintepretasikan informasi demografi pelanggan sebagai hasil identifikasi sistem. Pada penelitian ini akan umum kamus data disebut juga sebagai *Gazeteer*. Adapun perancangan kamus data dilakukan dengan mengambil data dari sumber yang telah ada pada penelitian-penelitian sebelumnya. Terdapat 3 perancangan kamus data yaitu gazeteer kewarganegraan, gazeteer profesi, dan gazeteer alamat.

4.3.1. Perancangan Gazeteer Kewarganegaraan dan Jenis Kelamin

Pada penelitian ini, implementasi perancangan gazeteer kewarganegaraan akan menghasilkan daftar nama-nama orang dan kewarganegaraannya. Untuk mengurangi bias mengenai korelasi antara nama pada jenis kewarganegaraan maka digunakan penelitian yang sudah ada pada sebelumnya. Implementasi perancangan gazetter pada penelitian ini menggunakan referensi sebuah website reporsitori nama *Behind the Name* [18].

Proses untuk mendapatkan dan memuat data gazeteer kewarganegaraan pada aplikasi GATE Developer melalui beberapa tahap. Tahap yang pertama adalah mengekstraksi data dengan mengakses web *Behind the Name*. Ekstraksi dilakukan secara manual yaitu dengan mengambil data satu-persatu. Tahapan selanjutnya adalah menyimpan data yang telah diambil dalam format .lst agar dapat digunakan aplikasi GATE. Setiap file .lst dibuat berdasarkan jenis kewarganegaraan nama, sehingga setiap file .lst mewakili informasi kewarganegaraan.

Tahapan ketiga yaitu tahapan akhir adalah memuat file .lst yang telah dibuat kedalam aplikasi GATE.



Gambar 4.5 Perancangan Gazeteer Nama

Gambar 4.5 diatas menggambarkan rancangan gazeteer nama. Dapat dilihat dari sumber data yang diekstrak dari situs *Behind the Name* menghasilkan 4 jenis data yang mencakup data *first name male*, data *surname male*, data *first name female*, dan data *surname female*.

Male Name
Dimitar Dobrev (First Name) (Surname)
Female Name
Diana Dobreva (First Name) (Surname)

Gambar 4.6 Contoh Data Nama

Gambar 4.6 diatas menjelaskan mengapa dibentuk 4 jenis data nama yang berbeda seperti yang disebutkan

sebelumnya. Terdapat *First Name* yang bersifat unisex yang berarti dapat diterapkan pada nama berjenis kelamin *male* maupun *female*. Namun juga terdapat nama yang bersifat unik seperti yang dicantumkan juga pada gambar 4.6, kondisi inilah yang menjadi dasar dibentuknya 4 jenis data yang berbeda. Kondisi ini juga berlaku dengan data *surname*.

Selain dibedakan berdasarkan jenis nama (*first name* dan *surname)* dan jenis kelamin, data juga dibedakan berdasarkan kewarganegaraannya. Sehingga bentuk penyimpanan data dipisah berdasarkan jenis nama, kelamin, dan kewarganegaraan.

4.3.2. Perancangan Gazeteer Profesi

Perancangan gazeteer profesi akan menghasilkan daftar profesi yang akan digunakan sebagai acuan data dalam mendeteksi pola yang memiliki ciri-ciri menggambarkan suatu profesi. Pada penelitian ini, pengambilan data untuk kebutuhan gazeteer profesi didapatkan dari situs data.gov yang dibawah kelola oleh *U.S. General Services Administration.* Data profesi akan disimpan dalam bentuk file .lst.

4.3.3. Perancangan Gazeteer Address

Perancangan gazeteer *address* akan menghasilkan daftar lokasi yang terbagi-bagi berdasarkan jenis lokasi. Gazeteer address akan digunakan sebagai acuan data dalam mendeteksi pola yang memiliki ciri-ciri menggambarkan suatu tempat atau lokasi.

Pada penelitian ini, pengambilan data untuk kebutuhan gazeteer *address* didapatkan dari situs *United Nations Economic Commission for Europe* (UNECE) [19]. Adapun daftar jenis lokasi pada penelitian ini yang terbagi menjadi berikut:

- Canton
- Capital City

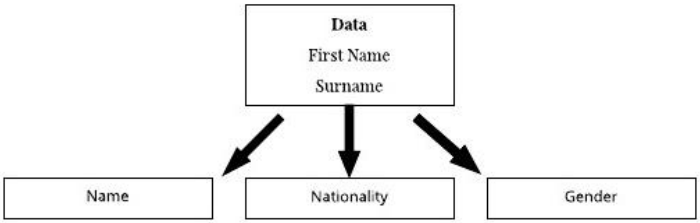
- Council Area
- County
- Department
- District
- Municipality
- Prefecture
- Province
- Region
- State
- Nation

Data lokasi akan disimpan dalam bentuk file .lst dan secara terpisah-pisah berdasarkan jenis lokasinya.

4.4. Perancangan Rule

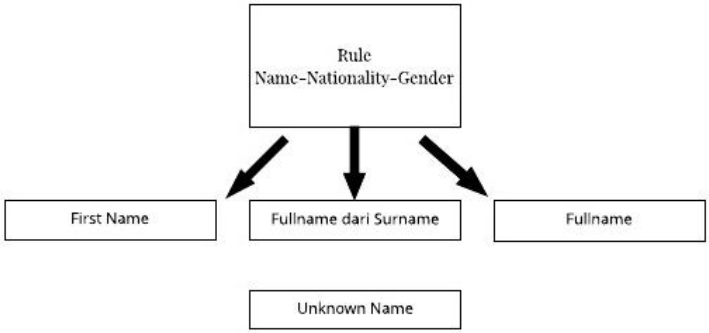
Tahapan perancangan rule mencakup hal-hal apa saja yang perlu dilakukan dalam menghasilkan sistem yang dapat mendeteksi atribut demografi pelanggan yang telah ditetapkan. Tahapan ini mencakup perancangan *tokenizer* yaitu memilah-milah input text menjadi perkata yang terpisah, *sentence splitter* yang merupakan bentuk *finite-state transducer* yang melakukan segmentasi dari teks utuh menjadi per-kalimat, POS *tagger* yang melakukan *tagging* di setiap kata, menjelaskan penggunaan *gazeteer* dalam mendefinisikan atribut demografi pelanggan. Perancangan rule pada penelitian ini menggunakan aplikasi GATE Developer sebagai lingkungan pengembangan. Secara garis besar terdapat 5 rule yang mencakup *name-nationality-gender, profession, address, phone,* dan email.

4.4.1. Perancangan Rule Name-Nationality-Gender



Gambar 4.7 Acuan Data Rule Name-Nationality-Gender

Perancangan rule *name-nationality-gender* menghasilkan 3 intepretasi sesuai dengan namanya. Ketiga intepretasi ini dijadikan satu sebab seluruhnya tergantung dengan data nama seperti yang diilustrasikan pada gambar 4.8.



Gambar 4.8 Perancangan Rule Name-Nationality-Gender

Pada rule ini terdapat 4 *grammar*, yaitu *grammar First Name, grammar Fullname* dari *Surname, grammar Fullname*, dan *grammar Unknown Name*. Pada tabel 4.4 menjelaskan gazeteer yang digunakan dan fungsi dari masing-masing *grammar.*

Tabel 4.4 Gazeeter dan Fungsi Grammar Nama

Grammar	Gazeteer	Fungsi
First Name	*Firstname*	Mendeteksi teks yang memiliki ciri-ciri menyerupai nama depan dengan menggunakan gazeteer *Firstname.*
Full Name dari *Surname*	*Surname*	Mendeteksi teks yang memiliki ciri-ciri menyerupai nama belakang/keluarga dan mengintepretasikannya sebagai nama panjang walaupun teks nama depan yang terdeteksi tidak terdapat pada gazeteer *Firstname*.
Fullname	*Firstname* dan *Surname*	Mendeteksi teks yang memiliki ciri-ciri menyerupai nama lengkap dengan mengkombinasikan gazeteer *Firstname* dan *Surname.*
Unknown Name	-	Mendeteksi teks yang memiliki ciri-ciri menyerupai nama tanpa menggunakan *gazeteer.*

4.4.1. Perancangan Rule Phone

Perancangan rule *Phone* berfungsi untuk mendeteksi pola yang memiliki ciri-ciri seperti nomor telefon. Terdapat

beberapa jenis format nomor telepon yang dijelaskan pada tabel 4.5 dibawah:

Tabel 4.5 Jenis Format Nomor Telepon

Jenis Nomor	Deskripsi	Contoh
Nomor Kode Negara	Pola nomor kode negara	+62
Nomor Kode Area	Pola nomor kode area	081
Nomor Regristrasi	Pola nomor regristrasi	8164474
Nomor Ekstensi	Pola nomor ekstensi	Ext. 1234/5
Prefix	Pola prefix	No Telp.

Pada terapannya terdapat berbagai kombinasi jenis nomor dalam membentuk kesatuan nomor telepon. Tabel 4.6 berikut mencakup fungsi apa saja yang perlu dibangun untuk mendeteksi berbagai format nomor telepon:

Tabel 4.6 Pola dan Fungsi Nomor Telepon

Fungsi	Jenis Nomor Yang Terlibat
Mendeteksi pola nomor telefon mencakup nomor kode area dan regristrasi.	• Nomor kode area • Nomor regristasi
Mendeteksi pola nomor telefon yang menggunakan prefix, kode area, dan regristrasi.	• Prefix • Nomor kode area • Nomor regristasi
Mendeteksi pola nomor telefon yang menggunakan kode negara	• Prefix • Nomor kode negara • Nomor regristrasi

Fungsi	Jenis Nomor Yang Terlibat
Mendeteksi pola nomor telefon yang hanya memiliki *phone extension*	• Nomor ekstensi
Mendeteksi pola nomor telefon yang memiliki angka *phone extension* yang berposisi di akhir nomor telefon.	• Nomor kode area • Nomor regristrasi • Nomor ekstensi
Mendeteksi pola nomor telefon yang memiliki kode area dan kode *phone extension*	• Prefix • Nomor kode area • Nomor regristrasi • Nomor ekstensi
Mendeteksi pola nomor telefon yang hanya mencakup prefix dan nomor regristasi telefon	• Prefix • Nomor Regristrasi
Mendeteksi pola nomor yang tidak umum	• Prefix

Tabel 4.7 berikut menjelaskan contoh dan pola jenis nomor yang diterapkan pada setiap grammar yang dibentuk pada tabel 4.6.

Tabel 4.7 Contoh dan Penjelasan Grammar Nomor Telepon

Grammar	Contoh dan Penjelasan
PhoneReg	0251 555555 (0251) 555555
	Nomor kode area + nomor regristrasi
PhoneRegContext	No telp: 0251 555555
	Prefix + nomor kode area + nomor regristrasi
PhoneFull	+62 321 32132 Tel: (+62) 321 32132

Grammar	Contoh dan Penjelasan
	Prefix + nomor kode negara + Nomor regristrasi
PhoneExt	Ext. 4444
	Nomor ekstensi
PhoneRegExt	0251 321321 ext. 4444
	Nomor kode area + nomor regristrasi + nomor ekstensi
PhoneRegExtContext	Tel: 0251 321321 ext. 4444
	Prefix + nomor kode area + nomor regristrasi + nomor ekstensi
PhoneNumberOnly	0812888888 Tel: 0812888888
	Prefix + nomor regristrasi
PhoneOtherContext	Tel: 123456789
	Prefix

4.4.2. Perancangan Rule Email

Perancangan rule email menghasilkan grammar yang bertugas dalam mendeteksi pola menyerupai alamat email. Berdasarkan RFC 5321 yaitu sebuah standar protokol internet yang memiliki spesifikasi khusus untuk *Internet Messaging Format* (IMF), alamat email terbagi menjadi 2 bagian yaitu *local part* dan *domain part*. *Local part* merupakan bagian yang umumnya dijadikan identitas dalam sebuah alamat email. Adapun spesifikasi *local part* yang valid dengan menggunakan beberapa karakter ASCII sebagai berikut:

- Huruf besar dan huruf kecil (A-Z, a-z) (ASCII: 65-90, 97-122).
- Digit (0 sampai 9) (ASCII: 48-57).
- Karakter !#$%&'*+-/=?^_`{|}~ (ASCII: 33, 35-39, 42, 43, 45, 47, 61, 63, 94-96, 123-126).
- Karakter . (titik) (ASCII: 46). Namun dengan penggunaan tidak pada diawal alamat email dan tidak

digunakan secara berurutan (Contoh: abc..def@mail.com tidak dibolehkan).

Bagian *domain part* mencakup nama domain yang umumnya mencantumkan nama dari penyedia layanan email. *Domain name* terdiri dari huruf, angka, tanda hubung, dan tanda titik. Maka dari itu, perancangan rule email mengacu pada jenis-jenis format email pada 4.8 berikut:

Tabel 4.8 Jenis-jenis Format Email

Jenis	Contoh
Dots	john.d@example.com j.doe@example.com john.doe@example.com
Underscores	john_d@example.com j_doe@example.com john_doe@example.com
Hyphens	john-d@example.com j-doe@example.com john-doe@example.com

Oleh karena itu perancangan rule email dibentuk sedemikian rupa untuk mendeteksi jenis-jenis format email yang tertera pada tabel 4.8.

4.4.1. Perancangan Rule Profesi

Perancangan rule profesi menghasilkan grammar yang bertugas dalam mendeteksi pola yang memiliki ciri-ciri menggambarkan suatu profesi. Perancangan rule profesi cukup sederhana yaitu cukup dengan menggunakan gazeteer profesi sebagai acuan dalam mendeteksi pola-pola profesi.

4.4.2. Perancangan Rule Address

Perancangan rule address menghasilkan grammar yang bertugas dalam mendeteksi pola yang memiliki ciri-ciri

menggambarkan suatu tempat atau lokasi. Berdasarkan format sumber data yang didapat dari United Nations Economic Commission for Europe (UNECE) [19], terdapat berbagai jenis lokasi yang berbeda-beda. Oleh karena itu perancangan rule dibentuk sedemikian rupa untuk dapat mendeteksi masing-masing jenis lokasi. Berikut ini merupakan daftar rule yang harus dibangun agar dapat mendeteksi masing-masing pola jenis lokasi:

- Rule Canton
- Rule Capital City
- Rule Council Area
- Rule County
- Rule Department
- Rule District
- Rule Municipality
- Rule Prefecture
- Rule Province
- Rule Region
- Rule State
- Rule Country

BAB V
IMPLEMENTASI

Bab ini berisi tentang proses implementasi dalam tugas akhir yang mencakup pengembangan rules dari setiap atribut demografi yang telah ditentukan dari hasil pemetaan.

5.1. Perangkat Penelitian

Pada pengembangan aplikasi, peneliti menggunakan perangkat keras dengan spesifikasi seperti pada Tabel 5.1. Sedangkan untuk perangkat lunak yang digunakan dalam pengembangan aplikasi adalah seperti pada Tabel 5.2.

Tabel 5.1 Perangkat keras implementasi penelitian

Perangkat keras	Spesifikasi
Laptop	Prosesor: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHz
	RAM: 8.00 GB

Tabel 5.2 Perangkat lunak implementasi penelitian

Perangkat lunak	Spesifikasi
Sistem Operasi	Windows 10
Web Server	Apache
Database	MySQL
Rule development environment	GATE Developer
Editor	Notepad++

5.2. Pengambilan Data Review Google Play

Pengambilan data merupakan proses awal yang dilakukan pada penelitian ini. Pengambilan data dilakukan melalui *web scraping* yang menggunakan cURL (client Uniform Resource Locator) untuk mengambil data dalam format json. Data json yang didapatkan selanjutnya dilakukan *decode* untuk melakukan ekstraksi nilai yang terdapat didalam json dan disimpan kedalam variabel sebagai wadah penyimpanan nilai json. Variabel yang sudah menampung nilai json selanjutnya dilakukan *query* yang befungsi memasukkan data kedalam basis data.

```
<?php

$db_user = 'root';
$db_pass = '';
$db_name = 'googleplay_api';

// connect to database
$dsn = 'mysql:dbname=' . $db_name . ';charset=utf8;host=localhost';
try {
    $dbh = new PDO( $dsn, $db_user, $db_pass );
} catch ( PDOException $e ) {
    echo 'Koneksi Gagal : ' . $e->getMessage();
}

?>
```

Gambar 5.1 Konfigurasi koneksi basis data

Gambar 5.1 diatas mencakup kode konfigurasi koneksi ke basis data googleplay_api.

5.3. Pembuatan *Scraper*

Pada bagian ini menjelaskan proses *scraper* yang bertugas dalam pengambilan data *review* app pada Google Play API. Bentuk data yang dilakukan *scrap* dalam bentuk Json sehingga menggunakal proses cURL (*client* URL).

```
include('koneksi.php');
for ($x = 0; $x <= 100; $x++) {
$query = "select id, appid from applications where id = $x";
$result1 = $dbh->query($query)->fetchAll();
foreach($result1 as $r1){
for ($m = 0; $m <= 100; $m++) {
$a =  $r1['appid'];
$mx = $r1['id'];
$service_url =
"http://gooplayapi.herokuapp.com/reviews/$a/en/newest/$m
";
$curl = curl_init($service_url);
curl_setopt($curl, CURLOPT_RETURNTRANSFER, 1);
curl_setopt($curl, CURLOPT_SSL_VERIFYHOST, 0);
curl_setopt($curl, CURLOPT_SSL_VERIFYPEER, 0);
$curl_response = curl_exec($curl);
curl_close($curl);
$curl_jason = json_decode($curl_response, true);
```

Gambar 5.2 Memasukkan data ke proses cURL

Pada gambar 5.2 merupakan kode yang bertugas dalam menjalankan proses cURL (*client*URL) untuk mendapatkan data dalam format json. Pada di line 3 dapat dilihat bahwa query dilakukan pada tabel applications untuk mendapatkan nilai appid, hal ini dilakukan agar program dapat melakukan *scraping* terhadap seluruh appid yang telah tersedia secara otomatis tanpa perlu mendefinisikan url yang perlu dilakukan scrap secara satu-persatu. Data json yang telah didapatkan kemudian dilakukan *decode*, dan untuk mengintepretasikan

nilai dari json yang telah dilakukan *decode* selanjutnya perlu dilakukan *looping*.

```
foreach ($curl_jason as $vali) {
   if (isset($vali['userName'])) {
         $userName = htmlspecialchars($vali['userName'],
ENT_QUOTES);
      } else {
         $userName = "";
      }
                  if (isset($vali['userImage'])) {
         $userImage = htmlspecialchars($vali['userImage'],
ENT_QUOTES);
      } else {
         $userImage = "";
      }
      if (isset($vali['date'])) {
         $date = htmlspecialchars($vali['date'],
ENT_QUOTES);
      } else {
         $date = "";
      }
      if (isset($vali['url'])) {
         $url = htmlspecialchars($vali['url'], ENT_QUOTES);
      } else {
         $url = "";
      }
      if (isset($vali['score'])) {
         $score = htmlspecialchars($vali['score'],
ENT_QUOTES);
      } else {
         $score = "";
      }
      if (isset($vali['title'])) {
         $title = htmlspecialchars($vali['title'],
ENT_QUOTES);
      } else {
```

```
        $title = "";
      }
      if (isset($vali['text'])) {
        $text = htmlspecialchars($vali['text'],
ENT_QUOTES);
      } else {
        $text = "";
      }
```

Gambar 5.3 Kode proses pengolahan data dari cURL

Kode pada Gambar 5.3 mencakup kode yang menjalankan tugas pengolahan data yang telah didapatkan dari cURL (*client* URL). Pada kode ini dilakukan konversi karakter data json yang sebelumnya belum terdefinisi menjadi entitas HTML dengan fungsi htmlspecialchars.

```
$query1 = "select * from reviews where url = '$url'";
  $result2 = $dbh->query($query1)->fetch();
  if($result2==false){
  $table = 'reviews';
  $field =
'`app_id`,`userName`,`userImage`,`date`,`url`,`score`,`title`,
`text`';
  $val = '?,?,?,?,?,?,?,?,?';
  $array = array($mx, $userName, $userImage, $date, $url,
$score, $title, $text );
  $sth = $dbh->prepare("INSERT INTO $table ($field)
VALUES ($val)");
  $input = $sth->execute($array);
  var_dump($input);
```

Gambar 5.4 Memasukkan data dari proses list

Kode pada gambar 5.4 mencakup kode yang menjalankan proses pengecekan data sehingga tercegah terjadinya redudansi pada database atau terdapat data yang

sama persis nilainya pada database. Pada kode diatas dapat dilihat cara pengecekannya yaitu dengan mencocokkan url yang merupakan *unique key*. Selanjutnya dijalankan proses pengambilan data *review* yang tersedia pada Google Play API yang mencakup app_id, userName, userImage, date, url, score, title, text. Seluruh data yang telah didapatkan selanjutnya dilakukan *insert* ke tabel *reviews*.

5.4. Praproses Data

Pada bagian ini menjelaskan bagaimana melakukan persiapan data yang merubah data menjadi siap untuk diunggah dan diproses pada aplikasi GATE Developer. Langkah pertama yang dilakukan pada tahapan ini adalah merubah karakter pemisah antar data dari titik koma (;) menjadi spasi. Perubahan dilakukan dengan merubah pengaturan *list separator* yang terdapat pada *control panel* sistem operasi.

Selanjutnya adalah dengan menghilangkan karakter kutip (“) dengan menggunakan *Visual Basic for Applications* (VBA).

```
Sub Export()
'updateby Extendoffice 20160530
    Dim xRg As Range
    Dim xRow As Range
    Dim xCell As Range
    Dim xStr As String
    Dim xTxt As String
    Dim xName As Variant
    On Error Resume Next
    If ActiveWindow.RangeSelection.Count >
1 Then
      xTxt =
ActiveWindow.RangeSelection.AddressLocal
    Else
      xTxt =
ActiveSheet.UsedRange.AddressLocal
    End If
```

```
    Set xRg = Application.InputBox("Please
select data range:", "Kutools for Excel",
xTxt, , , , , 8)
    If xRg Is Nothing Then Exit Sub
    xName =
Application.GetSaveAsFilename("", "CSV File
(*.csv), *.csv")
    Open xName For Output As #1
    For Each xRow In xRg.Rows
        xStr = ""
        For Each xCell In xRow.Cells
            xStr = xStr & xCell.Value &
Chr(9)
        Next
        While Right(xStr, 1) = Chr(9)
            xStr = Left(xStr, Len(xStr) - 1)
        Wend
        Print #1, xStr
    Next
    Close #1
    If Err = 0 Then MsgBox "The file has
saved to: " & xName, vbInformation, "Kutools
for Excel"
End Sub
```

Gambar 5.5 Kode VBA Menghilangkan Tanda Kutip

Gambar 5.5 diatas mencakup kode VBA yang digunakan dalam menghilangkan karakter kutip (“) pada file *Comma Separated Value* (.CSV). Pada Gamber 5.6 berikut ini merupakan gambaran hasil akhir dari praproses data.

2 1 Karen Powers https://lh5.googleusercontent.com/-pVqcdWLIhOk/AAAAAAAAAAI/AAAAAAAAAAA/AHalGhog4nQaAiXSRFv3u_5Fd5XkV 5/21/2017 https://play.google.com/store/apps/details?id=com.netflix.mediaclient&reviewId=Z3A6QU9xcFRPSFR4VXpBa29Ma1FCX202R0lQMEs0bC1EbEp2bVphenowLXZReUZjZTZyUzdiWkRLV3VuQm5oVF9FanFfbDFCSnU2OXNVa29SMFNGTGJ3R 5 Netflix rocks!

Gambar 5.6 Data Setelah Diproses

5.5. Pemetaan Ketersediaan Data Dalam Identifikasi Demografi

Pada bagian ini akan dijelaskan proses pemetaan ketersediaan data untuk melihat seberapa jauh potensi identifikasi demografi pelanggan dengan mencocokkan variabel demografi pelanggan yang telah ditetapkan yang dapat dilihat pada tabel 4.3. Pada tabel 5.3 dibawah menjelaskan ketersediaan data yang disediakan oleh Google Play API:

Tabel 5.3 Ketersediaan data

Nama Atribut	**Keterangan**
ID	Merupakan ID dari user Google Play Store
userName	Teks dari user name pereview aplikasi Google Play
userImage	Gambar *profile picture* dari pereview aplikasi
date	Tanggal dari komentar
url	URL dari komentar pereview
Score	Nilai yang diberikan oleh pereview terhadap aplikasi
Title	Judul dari komentar pereview

Nama Atribut	Keterangan
Text	Isi dari komentar pereview terhadap aplikasi

Berdasarkan data yang tersedia dilakukan identifikasi variabel demografi pelanggan apa saja yang dapat diintepretasikan.

Tabel 5.4 Potensi identifikasi dari data yang tersedia

Nama Atribut	Potensi Identifikasi
ID	• Tidak dapat digunakan
userName	• Nationality • Gender
userImage	• Umur
date	• Tidak dapat digunakan
url	• Tidak dapat digunakan
Score	• Tidak dapat digunakan
Title	• Tidak dapat digunakan
Text	• Bahasa yang digunakan • Profesi

Pada tabel 5.5 dibawah mencakup hasil pemetaan sementara, yaitu pemetaan antara ketersediaan data dengan atribut demografi pelanggan yang belum dilakukan pengujian untuk melihat apakah data yang tersedia dapat diidentifikasi menjadi data demografi pelanggan sesuai atribut yang telah ditetapkan pada tabel 4.3.

Tabel 5.5 Hasil pemetaan

Atribut	**Keterangan**
Identity	Data *userName* dapat dijadikan identitas. Data *userName* dapat menjelaskan kewarganegaraan.
Gender	Dapat diidentifikasi dengan menggunakan data *userName.*
Age	-
Profession	Dapat diidentifikasi dengan menggunakan data *Comment Text.*
Email	Dapat diidentifikasi dengan membangun fungsi Regex yang mendeteksi adanya format email pada data *Comment Text.*
Address	Dapat diidentifikasi dengan menggunakan data *Comment Text.*
Phone	Dapat diidentifikasi dengan membangun fungsi Regex yang mendeteksi adanya pola nomor telefon pada data *Comment Text.*
Product_preference	-
User_rank	-

Berdasarkan hasil pemetaan, didapatkan model template demografi pelanggan yang akan digunakan pada penelitian dengan daftar atribut sebagai berikut:

Tabel 5.6 Template Demografi Pelanggan

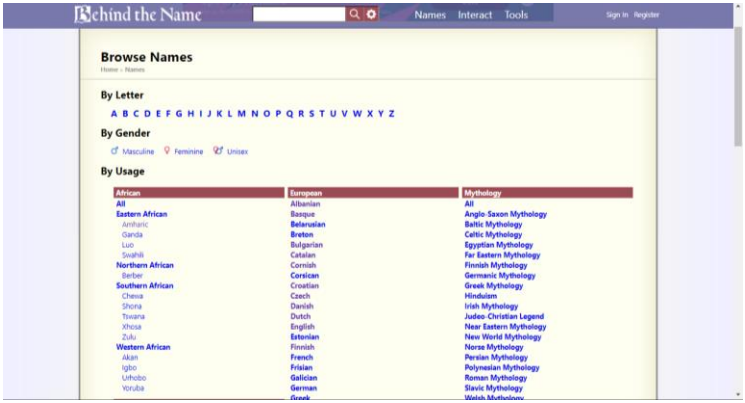
Nationality
Gender
Profession
Email
Address
Phone

5.6. Implementasi Gazeteer

Tahapan ini merupakan implementasi dari tahapan analisis dan perancangan *gazeteer*. Sesuai dengan yang dijelaskan pada perancangan *gazeteer*, terdapat 3 jenis perancangan kamus data yaitu gazeteer kewarganegaraan dan jenis kelamin, gazeteer profesi, dan gazeteer alamat.

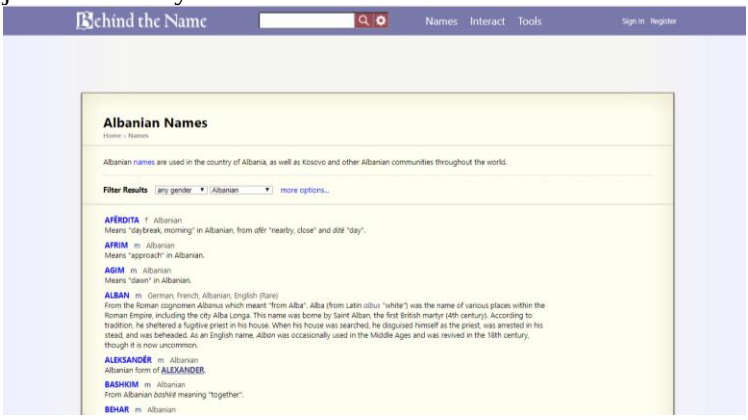
5.6.1. Implementasi Gazeteer Kewarganegaraan dan Jenis Kelamin

Data nama, kewarganegaraan dan jenis kelamin didapatkan melalui situs *behindthenames*. Pada gambar 5.7 menunjukkan daftar negara yang didalamnya terdapat data nama dan jenis kelamin.



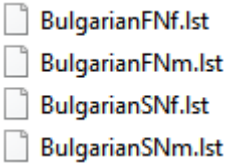
Gambar 5.7 Tampilan Situs Behind The Name

Pada gambar 5.8 berikut menunjukkan isi dari tautan negara seperti yang ditunjukkan pada gambar 5.7. Isinya mencakup daftar nama dan dapat diterapkan filter berdasarkan jenis kelaminnya.



Gambar 5.8 Daftar Nama

Selanjutnya dilakukan ekstraksi dari setiap daftar yang ada dan disimpan kedalam file berformat .lst. File .lst disimpan terpisah berdasarkan jenis kewarganegaraannya dan jenis kelamin seperti ditunjukkan pada gambar 5.9.



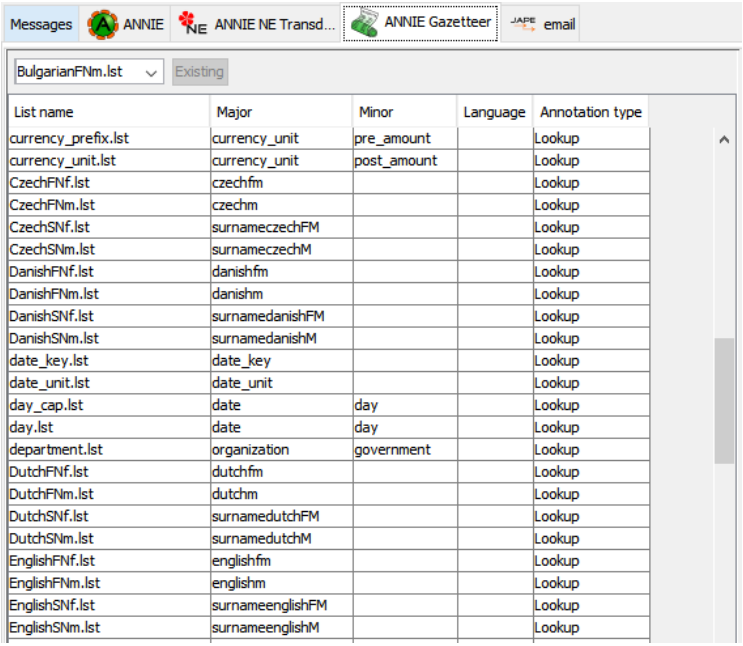
Gambar 5.9 Penyimpanan File Gazeteer

Tabel 5.7 berikut merupakan tabel yang mencakup keterangan dari penyimpanan file gazeteer.

Tabel 5.7 Keterangan Penyimpanan File Gazeteer

BulgarianFNm.lst	Negara: Bulgarian Jenis Data: FN (First Name) Jenis Kelamin: Male
BulgarianFNf.lst	Negara: Bulgarian Jenis Data: FN (First Name) Jenis Kelamin: Female
BulgarianSNm.lst	Negara: Bulgarian Jenis Data: SN (Surname) Jenis Kelamin: Male
BulgarianSNf.lst	Negara: Bulgarian Jenis Data: SN (Surname) Jenis Kelamin: Female

Tahapan berikutnya adalam memuat setiap file gazeteer kedalam lingkungan pemrograman GATE Developer. Cara mengoperasikannya cukup dengan memasukkan nama dari file gazeteer yang telah dibuat dan menekan tombol *"add"*.

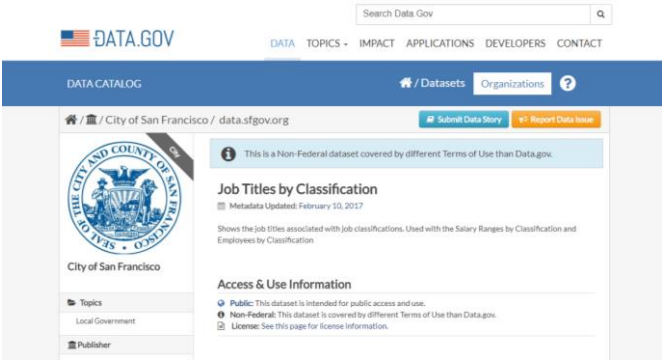


List name	Major	Minor	Language	Annotation type
currency_prefix.lst	currency_unit	pre_amount		Lookup
currency_unit.lst	currency_unit	post_amount		Lookup
CzechFNf.lst	czechfm			Lookup
CzechFNm.lst	czechm			Lookup
CzechSNf.lst	surnameczechFM			Lookup
CzechSNm.lst	surnameczechM			Lookup
DanishFNf.lst	danishfm			Lookup
DanishFNm.lst	danishm			Lookup
DanishSNf.lst	surnamedanishFM			Lookup
DanishSNm.lst	surnamedanishM			Lookup
date_key.lst	date_key			Lookup
date_unit.lst	date_unit			Lookup
day_cap.lst	date	day		Lookup
day.lst	date	day		Lookup
department.lst	organization	government		Lookup
DutchFNf.lst	dutchfm			Lookup
DutchFNm.lst	dutchm			Lookup
DutchSNf.lst	surnamedutchFM			Lookup
DutchSNm.lst	surnamedutchM			Lookup
EnglishFNf.lst	englishfm			Lookup
EnglishFNm.lst	englishm			Lookup
EnglishSNf.lst	surnameenglishFM			Lookup
EnglishSNm.lst	surnameenglishM			Lookup

Gambar 5.10 Memuat File Gazeteer Pada GATE

5.6.2. Implementasi Gazeteer Profesi

Data untuk kebutuhan gazeteer profesi didapatkan dari situs data.gov yang dibawah kelola oleh *U.S. General Services Administration*. Pada gambar 5.11 berikut merupakan tampilan halaman untuk mengunduh data profesi.



Gambar 5.11 Tampilan Situs data.gov

Selanjutnya yang dilakukan adalah mengunduh file berisikan data profesi dalam format .csv yang disediakan dalam halaman tersebut. File berisikan daftar nama profesi dan kode profesi dari profesi itu sendiri.

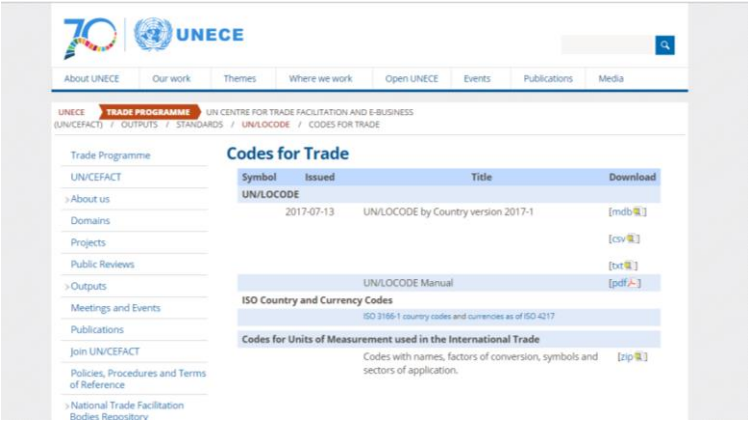
```
Job Code,Job Title
0111,"BdComm Mbr, Grp2,M=$25/Mtg"
0112,"BdComm Mbr, Grp3,M=$50/Mtg"
0114,"BdComm Mbr, Grp5,M$100/Mo"
0115,"BdComm MbrGrp6,D=@$100/Halfday"
0116,"Brd Comm Mbr, M=$200/Mtg"
0118,"BdComm Mbr, Grp7,M=$500/Month"
0130,Superintendent of Schools
0140,"Chief, Fire Department"
0150,Dep Chf of Dept (Fire Dept)
0170,Assistant Law Librarian
0180,Law Librarian
0190,Bookbinder
0382,Inspector 3
0390,Chief of Police
0395,Assistant Chief of Police
0402,Deputy Chief 3
0490,Commander 3
0720,"Member, Board of Supervisors"
0882,Mayoral Staff II
0884,Mayoral Staff IV
```

Gambar 5.12 Data Profesi data.gov

Selanjutnya dilakukan praproses data dengan meghilangkan data-data yang tidak dibutuhkan yaitu data kode profesi. Kemudian data disimpan dalam file berformat .lst.

5.6.3. Implementasi Gazeteer Address

Data untuk kebutuhan gazeteer address didapatkan dari situs United Nations Economic Commission for Europe (UNECE). Pada gambar 5.13 menunjukkan halaman untuk mengunduh data address.



Gambar 5.13 Tampilan Situs UNECE

Selanjutnya yang dilakukan adalah mengunduh file dalam format .csv yang disediakan pada halaman tersebut. File tersebut berisikan daftar address yang telah diberikan label jenis dari address itu sendiri.

```
77  "AM","SU","Syunik'","Region"
78  "AM","TV","Tavuš","Region"
79  "AM","VD","Vayoc Jor","Region"
80  "AO","BGO","Bengo","Province"
81  "AO","BGU","Benguela","Province"
82  "AO","BIE","Bié","Province"
83  "AO","CAB","Cabinda","Province"
84  "AO","CCU","Kuando Kubango","Province"
85  "AO","CNN","Cunene","Province"
86  "AO","CNO","Kwanza Norte","Province"
87  "AO","CUS","Kwanza Sul","Province"
88  "AO","HUA","Huambo","Province"
89  "AO","HUI","Huíla","Province"
90  "AO","LNO","Lunda Norte","Province"
91  "AO","LSU","Lunda Sul","Province"
92  "AO","LUA","Luanda","Province"
93  "AO","MAL","Malange","Province"
```

Gambar 5.14 Data Address UNECE

Selanjutnya dilakukan praproses data dengan menghilangkan data-data yang tidak dibutuhkan seperti inisial negara dan inisial nama address/daerah. Data yang diambil kemudian disimpan secara terpisah-pisah berdasarkan jenis daerah.

5.7. Implementasi Rule

Tahapan ini merupakan implementasi dari tahapan analisis dan perancangan rule ke dalam lingkungan pemrograman GATE Developer sesuai dengan penjelasan lingkungan implementasi.

5.7.1. Implementasi Rule Name-Nationality-Gender

Rule name-nationality-gender berperan sebagai *rule* yang bertugas dalam mendeteksi pola nama dan mengintepretasikan jenis kewarganegaraan dan jenis kelamin berdasarkan nama yang terdeteksi. Adapun pengembangan kode rule *name-nationality-gender* yang dibagi berdasarkan

gazeteer yaitu jenis kelamin, kewarganegaraan, dan jenis nama (*firstname, surname,* atau *fullname).*

No	Rule Name-nationality-gender
1	Phase: namenationalitygender
2	Input: Lookup Token
3	Options: control = Appelt debug = false
4	
5	//ARABIC MALE
6	Rule: arabicmale
7	Priority: 25
8	(
9	{Token.string == "1"}
10	)
11	(
12	{Lookup.majorType == arabicm}
13	)
14	:arabicm
15	-->
16	{
17	gate.AnnotationSet arabicm = (gate.AnnotationSet)bindings.get("arabicm");
18	gate.Annotation arabicmAnn = (gate.Annotation)arabicm.iterator().next();
19	gate.FeatureMap features = Factory.newFeatureMap();
20	features.put("Gazeteer", arabicmAnn.getFeatures().get("majorType"));
21	features.put("rule","arabicmale");
22	features.put("Type","First Name");
23	features.put("Gender","Male");
24	features.put("Nationality","Arabic");
25	outputAS.add(arabicm.firstNode(), arabicm.lastNode(), "ArabicMale",features);
26	}
27	

```
28  Rule: surnamearabicm
29  Priority: 25
30  (
31  {Token.string == "1"}
32  )
33  (
34  {Token.kind==word,Token.category==NNP,
    Token.orth==upperInitial}
35   (
36    {Lookup.majorType == surnamearabicM}
37   )
38  )
39  :surnamearabicm
40  -->
41  {
42  gate.AnnotationSet surnamearabicm =
    (gate.AnnotationSet)bindings.get("surn
    amearabicm");
43  gate.Annotation surnamearabicmAnn =
    (gate.Annotation)surnamearabicm.iterat
    or().next();
44  gate.FeatureMap features =
    Factory.newFeatureMap();
45  features.put("Gazeteer",
    surnamearabicmAnn.getFeatures().get("m
    ajorType"));
46  features.put("rule","surnamearabicm");
47  features.put("Type","Surname");
48  features.put("Gender","Male");
49  features.put("Nationality","Arabic");
50  outputAS.add(surnamearabicm.firstNode(
    ), surnamearabicm.lastNode(), "Arabic
    Fullname Male (Surname)",features);
51  }
52
53  Rule: fullnamearabicm
54  Priority: 50
55  (
56  {Token.string == "1"}
57  )
```

```
58  (
59   {Lookup.majorType == arabicm}
60   (
61   {Lookup.majorType == surnamearabicM}
62   )
63  )
64  :fullnamearabicm
65  -->
66  {
67  gate.AnnotationSet fullnamearabicm =
    (gate.AnnotationSet)bindings.get("full
    namearabicm");
68  gate.Annotation fullnamearabicmAnn =
    (gate.Annotation)fullnamearabicm.itera
    tor().next();
69  gate.FeatureMap features =
    Factory.newFeatureMap();
70  features.put("Gazeteer 1",
    fullnamearabicmAnn.getFeatures().get("
    majorType"));
71  features.put("Gazeteer
    2","surnamearabicM");
72  features.put("rule","fullnamearabicm")
    ;
73  features.put("Type","Full Name");
74  features.put("Gender","Male");
75  features.put("Nationality","Arabic");
```

Gambar 5.15 Rule Name-Nationality-Gender

Rule diatas merupakan salah satu rule dari sekian banyak negara. Berikut ini merupakan penjelasan dari rule:

- Line 1: Sebuah *Phase* terdiri dari satu atau beberapa rule atau pola. Dikarenakan rule yang dibuat adalah rule mengenai *name-nationality-gender* maka dinamakan `namenationalitygender`.
- Line 2: Jenis anotasi input yang digunakan pada rule adalah *Lookup* dan *Token*. Digunakan jenis anotasi input *Lookup* agar rule menggunakan gazeteer dan mencari teks yang

terdapat pada gazeteer. Selanjutnya digunakan anotasi input *Token* agar rule dapat memproses *token* yang akan diintepretasikan secara khusus oleh rule.

- Line 3: Jenis *Control* yang digunakan adalah *Appelt* dengan konfigurasi debug = false. Penggunaan jenis *control Appelt* memungkinkan rule yang dijalankan hanya satu pada pendeteksian teks tertentu walaupun terdapat kemungkinan lebih dari 1 rule lainnya yang juga mendeteksi teks tersebut, caranya adalah dengan menggunakan sistem prioritas.
- Rule First Name
    - Line 6: Penamaan dari setiap rule yang dibuat. Pada contoh diatas rule diberi nama *arabicmale.*
    - Line 7: Tingkat prioritas dari rule terkait. Semakin kecil nilainya semakin diprioritaskan.
    - Line 9: Rule akan membaca token input "1" terlebih dahulu sebelum membaca pola berikutnya.
    - Line 12: Setelah membaca input token selanjutnya rule akan menggunakan gazeteer.
    - Line 20: Intepretasi gazeteer yang terdeteksi dan digunakan pada teks.
    - Line 21: Intepretasi nama rule yang diterapkan.
    - Line 22: Intepretasi jenis teks yang terdeteksi.
    - Line 23: Intepretasi jenis kelamin dari nama yang terdeteksi.
    - Line 24: Intepretasi jenis kewarganegaraan dari nama yang terdeteksi.
    - Line 25: Nama anotasi yang akan muncul apabila teks terdeteksi.
- Rule Surname
    - Line 31: Rule akan membaca token input "1" terlebih dahulu sebelum membaca pola berikutnya.
    - Line 34: Rule akan membaca token dengan POS berjenis *word,* kategori token NNP (*noun singular*), dan dengan huruf awal *uppercase*.
    - Line 36: Rule akan menggunakan gazeteer surname yang tertera.
- Rule Fullname

- Line 56: Rule akan membaca token input “1” terlebih dahulu sebelum membaca pola berikutnya.
- Line 59: Rule akan menggunakan gazeteer surname yang tertera (*first name)*.
- Line 61: Rule akan menggunakan gazeteer surname yang tertera (*surname)*.

5.7.1. Implementasi Rule Phone

Rule phone berperan sebagai *rule* yang bertugas dalam mendeteksi pola nomor telepon dan mengintepretasikan jenis nomor telepon yang terdapat pada input. Seperti yang disebutkan pada tahapan perancangan, terdapat beberapa kombinasi jenis nomor telepon sehingga diperlukan kombinasi rule untuk mendeteksi masing-masing jenis nomor. Adapun pengembangan kode *rule phone* yang dibagi berdasarkan jenisnya seperti yang dijelaskan pada tabel 5.8 berikut:

Tabel 5.8 Daftar Rule Phone

PhoneReg
PhoneRegContext
PhoneFull
PhoneExt
PhoneRegExt
PhoneRegExtContext
PhoneNumberOnly
PhoneOtherContext

Berikut ini merupakan penjelasan dari setiap rule yang dibangun.

Tabel 5.9 Deskripsi Rule Phone

Grammar	Fungsi	Jenis Nomor Yang Terlibat
PhoneReg	Mendeteksi pola nomor telefon mencakup nomor kode area dan regristrasi.	• Nomor kode area • Nomor regristasi
PhoneRegContext	Mendeteksi pola nomor telefon yang menggunakan prefix, kode area, dan regristrasi.	• Prefix • Nomor kode area • Nomor regristasi
PhoneFull	Mendeteksi pola nomor telefon yang menggunakan kode negara	• Prefix • Nomor kode negara • Nomor regristrasi
PhoneExt	Mendeteksi pola nomor telefon yang hanya memiliki *phone extension*	• Nomor ekstensi
PhoneRegExt	Mendeteksi pola nomor telefon yang memiliki angka *phone extension* yang berposisi di akhir nomor telefon.	• Nomor kode area • Nomor regristrasi • Nomor ekstensi

Grammar	Fungsi	Jenis Nomor Yang Terlibat
PhoneRegExtContext	Mendeteksi pola nomor telefon yang memiliki kode area dan kode *phone extension*	• Prefix • Nomor kode area • Nomor regristrasi • Nomor ekstensi
PhoneNumberOnly	Mendeteksi pola nomor telefon yang hanya mencakup prefix dan nomor regristasi telefon	• Prefix • Nomor Regristrasi
PhoneOtherContext	Mendeteksi pola nomor yang tidak umum	• Prefix

Berikut ini merupakan rule phone yang dibangun untuk mendeteksi pola nomor telepon.

No	Rule Phone
1	`Phase: Phonenumber`
2	`Input:  Token Lookup`
3	`Options: control = appelt`
4	
5	`Macro: countrycode`
6	
7	
8	`({Token.string == "+"}`
9	`{Token.kind == number,Token.length == "2"}`
10	`)`
11	
12	`Macro: areacode`

```
13
14
15
16
17  (
      ({Token.kind == number,Token.length
18  == "3"} |
        {Token.kind == number,Token.length
19  == "4"} |
        {Token.kind == number,Token.length
20  == "5"} |
        {Token.kind == number,Token.length
21  == "6"} |
        {Token.kind == number,Token.length
22  == "7"} |
        {Token.kind == number,Token.length
23  == "8"} |
        {Token.kind == number,Token.length
24  == "9"} |
        {Token.kind == number,Token.length
25  == "10"} |
        {Token.kind == number,Token.length
26  == "11"} |
        {Token.kind == number,Token.length
27  == "12"})
28  |
29    ({Token.string == "("}
        ({Token.kind == number,Token.length
30  == "3"} |
         {Token.kind == number,Token.length
31  == "4"} |
         {Token.kind == number,Token.length
32  == "5"})
33    ({Token.string == ")"})
34   )
35  )
36
37  Macro: mobile
38
39
```

```
40
41  (
      ({Token.kind == number,Token.length
42  == "3"}
43
      ({Token.kind == number,Token.length
44  == "3"} |
      {Token.kind == number,Token.length
45  == "4"} |
      {Token.kind == number,Token.length
46  == "5"} |
      {Token.kind == number,Token.length
47  == "6"} |
      {Token.kind == number,Token.length
48  == "7"} |
      {Token.kind == number,Token.length
49  == "8"} |
      {Token.kind == number,Token.length
50  == "9"} |
      {Token.kind == number,Token.length
51  == "10"} |
      {Token.kind == number,Token.length
52  == "11"} |
      {Token.kind == number,Token.length
53  == "12"} |
      {Token.kind == number,Token.length
54  == "13"} |
       {Token.kind == number,Token.length
55  == "14"})
56    )
57  |
58
      ({Token.kind == number,Token.length
59  == "5"}|
      {Token.kind == number,Token.length
60  == "6"}|
      {Token.kind == number,Token.length
61  == "7"}|
```

```
   {Token.kind == number,Token.length
62 == "8"}|
   {Token.kind == number,Token.length
63 == "9"}|
   {Token.kind == number,Token.length
64 == "10"}|
   {Token.kind == number,Token.length
65 == "11"}|
    {Token.kind == number,Token.length
66 == "12"})
67 )
68
69 Macro: externalphone
70
71
72
73 (
74  (({Token.string == "x"})
   ({Token.string == "x"}{Token.string
75 == "."}) |
76   ({Token.string == "ext"})
   ({Token.string ==
77 "ext"}{Token.string == "."})
78  )
  ({Token.kind == number, Token.length
79 == "4"}|
   {Token.kind == number, Token.length
80 == "5"}|
   {Token.kind == number, Token.length
81 == "6"})
  ({Token.string == "/"}{Token.kind ==
82 number})?
83 )
84
85 Macro: PHONE_PREFIX
86 //Tel:
87
88 (
89  {Lookup.majorType == phone_prefix}
90  ({Lookup.majorType == phone_prefix})?
```

```
 91   ({Token.string == ":"})?
 92  )
 93
 94  ////////////////////////////////////
     ////////////
 95  Rule:PhoneNumber
 96  Priority: 20
 97
 98
 99
100  (
101   (areacode)
102   (mobile)
103  )
104  :phoneNumber -->
105   :phoneNumber.TelephoneNumber = {kind
     = "Telephone Register Number", rule =
     "PhoneNumber"}
106
107  Rule: PhoneRegContext
108  Priority: 100
109  // tel: 0114 222 1929
110  (
111   (PHONE_PREFIX)
112  )
113  (
114   (areacode)
115   (mobile)
116  )
117  :phoneNumber -->
118   :phoneNumber.Phone = {kind =
     "phoneNumber", rule =
     "PhoneRegContext"}
119
120  Rule:PhoneFull
121  Priority: 50
122
123
124  (
125   (PHONE_PREFIX)?
```

```
126 )
127 (
128  ((countrycode) |
129   ({Token.string == "("}
130    (countrycode)
131    {Token.string == ")"})
132  )
133
134  ({Token.string == "("}
135    {Token.string == "0"}
136    {Token.string == ")"})?
137
138  ({Token.kind == number,Token.length
     == "3"}|
139    {Token.kind == number,Token.length
     == "4"})
140
141  (mobile)
142 )
143 :phoneNumber -->
144  :phoneNumber.Phone = {kind =
     "phoneNumber", rule = "PhoneFull"}
145
146 Rule:PhoneExt
147 Priority: 20
148
149 (
150  (externalphone)
151 )
152 :phoneNumber -->
153  :phoneNumber.Phone = {kind =
     "phoneNumber", rule = "PhoneExt"}
154
155 Rule:PhoneRegExt
156 Priority: 40
157
158 (
159
160  (areacode)?
161  (mobile)
```

```
 (externalphone)
)
:phoneNumber -->
 :phoneNumber.Phone = {kind =
"phoneNumber", rule = "PhoneRegExt"}

Rule:PhoneRegExtContext
Priority: 40

(
 (PHONE_PREFIX)
)
(
 (areacode)?
 (mobile)
 (externalphone)
)
:phoneNumber -->
 :phoneNumber.Phone = {kind =
"phoneNumber", rule =
"PhoneRegExtContext"}

Rule:PhoneNumberOnly
Priority: 20

(PHONE_PREFIX)
(
 (mobile)
)
:phoneNumber -->
 :phoneNumber.Phone = {kind =
"phoneNumber", rule =
"PhoneNumberOnly"}
```

Gambar 5.16 Rule Phone

5.7.1. **Implementasi Rule Email**

Rule email berperan sebagai *rule* yang bertugas dalam mendeteksi dan mengintepretasikan pola email pada input teks.

Seperti yang dijelaskan pada bab perancangan, rule email dibentuk sedemikian rupa untuk mendeteksi jenis-jenis format email yang tertera pada tabel 4.8.

No	Rule Email
1	Phase: Email
2	Input: Token Lookup SpaceToken
3	Options: control = appelt
4	
5	Rule:Email
6	Priority: 50
7	(
8	(
9	{Token.kind == word}|
10	{Token.kind == number}
11	)[1,9]
12	(
13	{Token.string == "_"}
14	)?
15	({Token.string == "."})?
16	({Token.kind == word}|
17	{Token.kind == number}|
18	{Token.string == "_"}
19	)[0,9]
20	
21	{Token.string == "@"}
22	(
23	{Token.kind == word}|
24	{Token.kind == symbol}|
25	{Token.kind == punctuation}|
26	{Token.kind == number}
27	)
28	({Token.string == "."})?
29	(
30	{Token.kind == word}|
31	{Token.kind == symbol}|
32	{Token.kind == punctuation}|
33	{Token.kind == number}
34	)[0,9]

```
35   ({Token.string == "."})?
36  (
37    {Token.kind == word}|
38    {Token.kind == symbol}|
39    {Token.kind == punctuation}|
40    {Token.kind == number}
41   )?
42  ({Token.string == "."})?
43   (
44    {Token.string == "."}
45    (
46     {Token.kind == word}|
47     {Token.kind == number}
48    )
49    ({Token.string == "."})?
50    (
51     {Token.kind == word}|
52     {Token.kind == number}
53    )?
54    ({Token.string == "."})?
55    (
56     {Token.kind == word}|
57     {Token.kind == number}
58    )?
59   )
60  )
61  :emailAddress -->
      :emailAddress.Email= {kind = "Email
62  Address", rule = "emailaddress"}
```

Gambar 5.17 Rule Email

5.7.1. Implementasi Rule Profesi

Rule profesi berperan sebagai *rule* yang bertugas dalam mendeteksi dan mengintepretasikan pola teks yang memiliki ciri-ciri seperti nama sebuah profesi. Pengembangan kode rule profesi cukup sederhana, yaitu cukup dengan mengandalkan gazeteer profesi sebagai acuan data dalam mendeteksi pola profesi.

No	Rule Phone
1	Phase:
2	Input: Lookup Token
3	Options: control = appelt
4	
5	Rule: Profession
6	(
7	{Lookup.majorType == profesi}
8	(
9	{Lookup.majorType == profesi}
10	)?
11	)
12	:jobtitle
13	-->
14	:jobtitle.Profession = {rule = "Profession"}

Gambar 5.18 Rule Profesi

5.7.2. Implementasi Rule Address

Rule address berperan sebagai *rule* yang bertugas dalam mendeteksi pola yang memiliki ciri-ciri menggambarkan suatu tempat atau lokasi dan mengintepretasikan jenis lokasi yang terdeteksi. Adapun pengembangan kode *rule* address yang secara garis besar terbagi menjadi 2 jenis, yaitu pengembangan kode rule dari jenis-jenis lokasi seperti pada gambar 5.19. Kode rule berikut mewakili seluruh kode rule jenis lokasi lainnya sebab pola kodenya identik, hanya dilakukan pengubahan minor.

No	Rule Phone
1	Phase:addressphy
2	Input: Lookup Token
3	Options: control = Appelt debug = false
4	

```
5  Rule: cantonaddress
6  Priority: 50
7  (
8   {Lookup.majorType == canton_address}
9  )
10 :canton
11 -->
12 {
13 gate.AnnotationSet canton = (gate.AnnotationSet)bindings.get("canton");
14 gate.Annotation cantonAnn = (gate.Annotation)canton.iterator().next();
15 gate.FeatureMap features = Factory.newFeatureMap();
16 features.put("Gazeteer", cantonAnn.getFeatures().get("majorType"));
17 features.put("Address Type","Canton");
18 features.put("Rule","cantonaddress");
19 outputAS.add(canton.firstNode(), canton.lastNode(), "Address Canton",features);
20 }
```

Gambar 5.19 Rule Address

Jenis yang kedua adalah kode rule yang berperan dalam mendeteksi lokasi dalam kondisi dimana data lokasi tidak tersedia pada gazeteer. Metode yang diterapkan adalah dengan membaca token awal yang berisi tulisan “at” atau “in” kemudian diikuti dengan token berjenis NP. Pada Gambar 5.20 berikut merupakan kode dari jenis ini.

```
No  Rule Phone
1   Rule: UnknownLocRegion
2   Priority: 100
3   (
4    {Token.string == "at"}|
```

```
 5  {Token.string == "in"}|
 6  {Token.string == "At"}|
 7  {Token.string == "In"}
 8  (
 9  {Token.kind==word,Token.category==PN,Token.orth==upperInitial}
10  )
11 )
12
13 :location
14 -->
15  :location.LocationUnknown = {rule = "UnknownLocRegion"}
```

Gambar 5.20 Rule Address - Unknown Location

Halaman ini sengaja dikosongkan

BAB VI
HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan hasil serta analisis terhadap hasil yang diperoleh dari proses implementasi yang telah dibahas pada bab sebelumnya.

6.1. Ekstrak Data Review Google Play

Dari hasil ekstraksi data review pada Google Play, didapatkan daftar data-data review yang mencakup ID, app-id, username, url user image, tanggal review, url review, judul review, dan teks review seperti yang ditunjukkan pada gambar 6.1.

"2";"1";"Karen Powers";"https://lh5.googleusercontent.com/-pVqcdWLIhOk/AAAAAAAAAAI/AAAAAAAAAAA/AHalGhog4nQaAiXSRFv3u_5Fd5XkV";"May 21, 2017";"https://play.google.com/store/apps/details?id=com.netflix.mediaclient&reviewId=Z3A6QU9xcFRPSFR4VXpBa29Ma1FCX202R0lQMEs0bC1EbEp2bVphenowLXZReUZjZTZyUzdiWkRLV3VuQm5oVF9FanFfbDFCSnU2OXNVa29SMFNGTGJ3R";"5";;"Netflix rocks!"

Gambar 6.1 Hasil Ekstraksi Data Review

Berdasarkan hasil ekstraksi data review Google Play dapat dilihat terdapat karakter *noise* seperti tanda kutip (“) dan titik koma (;) yang akan mengganggu proses *tokenizing*. Untuk menghilangkan karakter *noise* tersebut diperlukan praproses data agar siap olah.

6.2. Praproses Data Review Google Play

Praproses data menghasilkan data review Google Play yang bersih dari karakter noise tanpa mengubah isi dari konten

data yang didapat. Gambar 6.2 merupakan hasil dari praproses data.

2 1 Karen Powers https://lh5.googleusercontent.com/-pVqcdWLIhOk/AAAAAAAAAAI/AAAAAAAAAAA/AHalGhog4nQaAiXSRFv3u_5Fd5XkV 5/21/2017 https://play.google.com/store/apps/details?id=com.netflix.mediaclient&reviewId=Z3A6QU9xcFRPSFR4VXpBa29Ma1FCX202R0lQMEs0bC1EbEp2bVphenowLXZReUZjZTZyUzdiWkRLV3VuQm5oVF9FanFfbDFCSnU2OXNVa29SMFNGTGJ3R 5 Netflix rocks!

Gambar 6.2 Hasil Praproses Data

6.3. Data Gazeteer

Data Gazeteer merupakan kamus data yang digunakan dalam melakukan pendeteksian pola-pola atribut demografi pelanggan. Terdapat 3 Gazeteer yang dibangun mencakup gazeteer kewarganegaraan dan jenis kelamin, gazeteer profesi, dan gazeteer address. Berikut merupakan penjelasan hasil dari masing-masing gazeteer yang dibangun.

6.3.1. Data Gazeteer Kewarganegaraan dan Jenis Kelamin

Dari hasil pengambilan data untuk kebutuhan gazeteer kewarganegaraan dan jenis kelamin, didapatkan sebanyak 14 negara dengan jumlah baris data pada masing-masing negara yang berbeda-beda. Pada tabel 6.1 berikut merupakan daftar negara yang dimuat pada penelitian:

Tabel 6.1 Daftar Negara

Arabic
Basque
Bulgarian
Catalan

Cornish
Croatian
Czech
Danish
Dutch
United States
Finnish
Indonesian
Thailand
Vietnam

Gazeteer negara dan jenis kelamin disimpan secara terpisah berdasarkan jenis kewarganegaraan, jenis nama, dan jenis kelaminnya seperti yang telah dijelaskan pada bab perancangan.

6.3.2. Data Gazeteer Profesi

Dari hasil pengambilan data untuk kebutuhan gazeteer profesi dari situs data.gov, didapatkan sebanyak 2464 data profesi dalam berbahasa Inggris. Pada tabel 6.2 berikut merupakan beberapa sampel dari gazeteer profesi.

Tabel 6.2 Sampel Data Gazeteer Profesi

cleric
clockmaker
clocksmith
coach
coachbuilder
coachman
coal miner
coalman

6.3.3. Data Gazeteer Address

Dari hasil pengambilan data untuk kebutuhan gazeteer address, terdapat 12 jenis address yang digunakan dalam penelitian ini. Berikut pada tabel 6.3 adalah daftar jenis daerah yang disimpan dan digunakan pada penelitian:

Tabel 6.3 Daftar Jenis Address

Canton
Capital City
Council Area
Country
County
Department
District
Municipality
Prefecture
Province
Region
State

Masing-masing dari jenis address memiliki jumlah address yang berbeda-beda. Berikut pada tabel 6.4 adalah sampel data dari jenis address Province:

Tabel 6.4 Sampel Data Address

Bayern
Bayrut
Bechar
Beja
Bejaia
Bekes
Bengkulu
Berat
Bayern

Bayrut
Bechar
Beja

6.4. Verifikasi Sistem

Verifikasi sistem dilakukan untuk melihat apakah metode penelitian dapat menjalankan tugasnya dalam mendeteksi pola-pola atribut demografi pelanggan beserta tingkat performa yang dihasilkan oleh masing-masing rule. Berikut ini merupakan verifikasi dari masing-masing rule yang telah dibangun. Pada tahapan ini dilakukan dengan memasukkan muatan data dengan masing-masing data yang telah diberi label kemudian diproses dengan sistem. Selanjutnya verifikasi dilakukan dengan membandingkan hasil anotasi dihasilkan oleh sistem yang bernilai benar sesuai dengan label muatan input data.

6.4.1. Muatan Data

Proses verifikasi pada masing-masing rule dilakukan dengan skenario muatan data input sebanyak 1000 data yang telah diberi label sesuai dengan jenis rule yang dibangun. Sebagai contoh 1000 data untuk rule Arabic Female, 1000 data untuk rule Arabic Female, dan seterusnya.

6.4.2. Verifikasi Rule Name-nationality-gender

Verifikasi rule *name-nationality-gender* dilakukan untuk mengetahui bagaimana performa rule ini dalam mengidentifikasi dan memberi anotasi input teks yang mengandung pola nama. Performa dari sebuah rule diukur dari tingkat akurasi anotasi yang bernilai benar sesuai dengan label input data.

Rule ini menghasilkan intepretasi berupa anotasi yang dilakukan berdasarkan hasil *tokenizing*, pengolahan data input dengan rule yang telah dibangun dan penyocokkan dengan data

gazeteer. Tingkat akurasi serta kesalahan anotasi juga dapat dilihat pada tabel 6.5.

Tabel 6.5 Akurasi dan Kesalahan Anotasi Nationality dan Gender

Arabic Female		
Tingkat Akurasi		97.6% (976 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Arabic Male	20	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Arabic Male* dan *Arabic Female*.
Unknown Name	4	Tidak terdapat data nama pada gazeteer *Arabic Female.*
Arabic Male		
Tingkat Akurasi		99.7% (997 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Unknown Name	3	Tidak terdapat data nama pada gazeteer *Arabic Female.*
Basque Female		
Tingkat Akurasi		96.5% (965 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**

Basque Male	34	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Basque Male* dan *Basque Female*.
Arabic Male	1	Terdapat data nama yang sama pada gazeteer *Basque Female* dengan *Arabic Male*.
Basque Male		
Tingkat Akurasi	100% (1000 dari 1000 input data)	
Bulgarian Female		
Tingkat Akurasi	98.8% (988 dari 1000 input data)	
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Bulgarian Male	12	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Bulgarian Male* dan *Bulgarian Female*.
Bulgarian Male		
Tingkat Akurasi	100% (1000 dari 1000 input data)	
Catalan Female		
Tingkat Akurasi	100% (1000 dari 1000 input data)	

Catalan Male		
Tingkat Akurasi		100% (1000 dari 1000 input data)
Cornish Female		
Tingkat Akurasi		96.7% (966 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Cornish Male	18	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Cornish Male* dan *Cornish Female*.
Unknown Name	15	Tidak terdapat data nama pada gazeteer *Cornish Female.*
Cornish Male		
Tingkat Akurasi		98.5% (985 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Unknown Name	15	Tidak terdapat data nama pada gazeteer *Cornish Male.*
Croatian Female		
Tingkat Akurasi		99% (990 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Croatian Male	10	Terdapat nama yang berjenis *unisex* sehingga terjadi

		redudansi pada data gazeteer *Croatian Male* dan *Croatian Female*.
Croatian Male		
Tingkat Akurasi		100% (1000 dari 1000 input data)
Czech Female		
Tingkat Akurasi		99.2% (992 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Croatian Female	8	Terdapat data nama yang sama pada gazeteer *Czech Female* dengan *Croatian Female*.
Czech Male		
Tingkat Akurasi		99.2% (992 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Croatian Male	8	Terdapat data nama yang sama pada gazeteer *Czech Male* dengan *Croatian Male*.
Danish Female		
Tingkat Akurasi		98.7% (987 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Danish Male	13	Terdapat nama yang berjenis *unisex*

		sehingga terjadi redudansi pada data gazeteer *Danish Male* dan *Danish Female*.
Danish Male		
Tingkat Akurasi		100% (1000 dari 1000 input data)
Dutch Female		
Tingkat Akurasi		97.2% (972 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Danish Female	5	Terdapat data nama yang sama pada gazeteer *Danish Female* dengan *Dutch Female*.
Indonesian Female	1	Terdapat data nama yang sama pada gazeteer Indonesian *Female* dengan *Dutch Female*.
Dutch Male	22	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Dutch Male* dan *Dutch Female*.
Dutch Male		
Tingkat Akurasi		99.7% (997 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**

Danish Male	3	Terdapat data nama yang sama pada gazeteer *Danish Male* dengan *Dutch Male*.
United States Female		
Tingkat Akurasi		99.4% (994 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Danish Female	5	Terdapat data nama yang sama pada gazeteer *Danish Female* dengan *United States Female*.
Basque Female	1	Terdapat data nama yang sama pada gazeteer *Basque Female* dengan *United States Female*.
United States Male		
Tingkat Akurasi		99.4% (994 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Danish Male	3	Terdapat data nama yang sama pada gazeteer *Danish Male* dengan *United States Male*.
Dutch Male	2	Terdapat data nama yang sama pada gazeteer *Dutch Male* dengan *United States Male*.

Indonesian Male	1	Terdapat data nama yang sama pada gazeteer *Indonesian Male* dengan *United States Male*.
Finnish Female		
Tingkat Akurasi	72.6% (726 dari 1000 input data)	
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
United States Female	247	Terdapat data nama yang sama pada gazeteer *United States Female* dengan *Finnish Female*.
United States Male	23	Terdapat data nama yang sama pada gazeteer *United States Male* dengan *Finnish Female*.
Finnish Male	4	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Finnish Male* dan *Finnish Female*.
Finnish Male		
Tingkat Akurasi	83.4% (834 dari 1000 input data)	
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
United States Male	147	Terdapat data nama yang sama pada gazeteer *United States*

		Male dengan *Finnish Male*.
United States Female	19	Terdapat data nama yang sama pada gazeteer *United States Female* dengan *Finnish Male*.
Indonesian Female		
Tingkat Akurasi		91.1% (911 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Indonesian Male	82	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Indonesian Male* dan *Indonesian Female*.
Arabic Female	5	Terdapat data nama yang sama pada gazeteer *Arabic Female* dengan *Indonesian Female*.
Unknown Name	2	Tidak terdapat data nama pada gazeteer *Indonesian Female.*
Indonesian Male		
Tingkat Akurasi		99.7% (997 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Arabic Male	1	Terdapat data nama yang sama pada gazeteer *Arabic Male*

		dengan *Indonesian Male*.
Unknown Name	2	Tidak terdapat data nama pada gazeteer *Indonesian Male*
Thailand Female		
Tingkat Akurasi		95.1% (951 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
Thailand Male	46	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Thailand Male* dan *Thailand Female*.
United States Male	1	Terdapat data nama yang sama pada gazeteer *United States Male* dengan *Thailand Female*.
United States Female	2	Terdapat data nama yang sama pada gazeteer *United States Female* dengan *Thailand Female*.
Thailand Male		
Tingkat Akurasi		99.8% (998 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
United States Male	2	Terdapat data nama yang sama pada gazeteer *United States*

		Male dengan *Thailand Male*.
Vietnam Female		
Tingkat Akurasi		74.2% (742 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
United States Female	61	Terdapat data nama yang sama pada gazeteer *United States Female* dengan *Vietnam Female*.
United States Male	190	Terdapat data nama yang sama pada gazeteer *United States Male* dengan *Vietnam Female*.
Vietnam Male	1	Terdapat nama yang berjenis *unisex* sehingga terjadi redudansi pada data gazeteer *Vietnam Male* dan *Vietnam Female*.
Unknown Name	7	Tidak terdapat data nama pada gazeteer *Vietnam Female.*
Vietnam Male		
Tingkat Akurasi		77.5% (775 dari 1000 input data)
Kesalahan Anotasi	**Jumlah Anotasi**	**Penyebab**
United States Male	208	Terdapat data nama yang sama pada gazeteer *United States*

		Male dengan *Vietnam Male*.
United States Female	10	Terdapat data nama yang sama pada gazeteer *United States Female* dengan *Thailand Male*.
Unknown Name	7	Tidak terdapat data nama pada gazeteer *Vietnam Male.*

Berdasarkan tabel 6.5 diatas dengan muatan data input teks sebanyak 1000 data, percobaan verifikasi yang dilakukan menghasilkan rata-rata tingkat akurasi mencapai 95.4%. Adapun tabel 6.6 yang menjelaskan peringkat akurasi setiap anotasi yang diurutkan dari yang tertinggi sebagai berikut.

Tabel 6.6 Urutan Rule Berdasarkan Tingkat Akurasi

Anotasi	Tingkat Akurasi
Basque Male	100%
Bulgarian Male	100%
Catalan Female	100%
Catalan Male	100%
Croatian Male	100%
Danish Male	100%
Thailand Male	99.8%
Arabic Male	99.7%
Indonesian Male	99.7%
Dutch Male	99.6%
United States Female	99.4%
United States Male	99.4%
Czech Female	99.2%
Czech Male	99.2%
Croatian Female	98.9%
Bulgarian Female	98.7%

Anotasi	Tingkat Akurasi
Danish Female	98.6%
Cornish Male	98.4%
Arabic Female	97.6%
Dutch Female	97.1%
Cornish Female	96.6%
Basque Female	96.5%
Thailand Female	95.1%
Indonesian Female	91.1%
Finnish Male	83.4%
Vietnam Male	77.4%
Vietnam Female	74%
Finnish Female	72.6%

Adapun penyebab bias rule *name-nationality-gender* yang secara garis besar disebabkan oleh adanya redudansi data pada gazeteer dan ketidaktersediaannya data nama pada gazeteer, berikut adalah penjelasan dari setiap bias yang muncul:

Tabel 6.7 Jenis Bias Rule Name-Nationality-Gender

Jenis Bias	Dampak
Nama berjenis Unisex sehingga terdapat pada gazeteer *male* dan *female*.	Terdapat kemungkinan sistem akan memberi anotasi jenis kelamin yang salah dengan jenis kewarganegaraan yang benar.
Terdapat nama di lebih dari 1 gazeteer negara	Terdapat kemungkinan sistem akan memberi anotasi kewarganegaraan yang salah.
Data nama tidak tersedia pada gazeteer	Sistem akan memberi label *Unknown Name* dimana tidak dijelaskan jenis kelamin dan kewarganegaraan dari nama yang terdeteksi.

Jenis Bias	Dampak
Data nama tidak tersedia pada gazeteer negara yang sesuai label input namun terdapat pada gazeteer negara lainnya.	Sistem akan memberi anotasi kewarganegaraan yang salah.

Berdasarkan percobaan verifikasi yang telah dilakukan, dapat ditemukan pola yang janggal yaitu adanya kesalahan anotasi jenis kelamin *unisex* yang selalu terjadi pada anotasi jenis kelamin *female*, tidak satupun kesalahan anotasi ini terjadi pada anotasi jenis kelamin *male*. Hal ini dikarenakan tingkat prioritas antar rule negara memiliki nilai yang sama, sedangkan cara kerja bahasa pemrograman JAPE yang akan menjalankan algoritma yang tercepat. Saat menjalankan sistem, sistem akan lebih dulu menjalankan algoritma jenis kelamin male sehingga apabila terdapat data nama yang bersifat unisex sistem akan memberi label sebagai jenis kelamin *male.* Ini juga akan mempengaruhi pada kesalahan anotasi apabila terdapat data nama di lebih dari satu gazeteer, sistem akan lebih dulu menjalankan algoritma yang tercepat sehingga adanya kekeliruan pemberian anotasi oleh sistem.

Hasil anotasi Vietnam dengan pengubahan prioritas rule/baris rule United States lebih awal

Type	Set	Start	End
Full Name Vietnam Female		2	12
Full Name Vietnam Female		30	38
Full Name Vietnam Female		60	68
Full Name Vietnam Female		74	84
Full Name Vietnam Female		90	97
Full Name Vietnam Female		103	111
Full Name Vietnam Female		129	138
Full Name Vietnam Female		156	167

740 Annotations (0 selected) Select:

Hasil anotasi Vietnam dengan pengubahan prioritas rule/ baris rule United States lebih akhir

Type	Set	Start	End
Full Name Vietnam Female		2	12
Full Name Vietnam Female		18	24
Full Name Vietnam Female		30	38
Full Name Vietnam Female		44	54
Full Name Vietnam Female		60	68
Full Name Vietnam Female		74	84
Full Name Vietnam Female		90	97
Full Name Vietnam Female		103	111

984 Annotations (0 selected) Select:

Gambar 6.3 Hasil Anotasi Dengan Perubahan Tingkat Prioritas

Gambar 6.3 diatas merupakan pembuktian dimana JAPE akan menjalankan algoritma tercepat. Ketika prioritas rule dirubah atau baris rule United States lebih awal daripada baris rule Vietnam sistem menghasilkan total anotasi yang tepat (Vietnam) sejumlah 740 anotasi. Sedangkan ketika baris rule United States dipindahkan lebih akhir daripada baris rule Vietnam menghasilkan total anotasi yang tepat (Vietnam) sejumlah 984 anotasi.

6.4.3. Verifikasi Rule Email

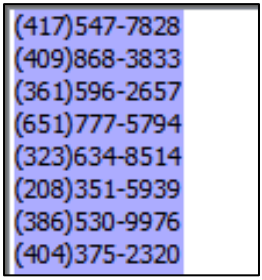
Verifikasi rule email dilakukan dengan melihat kemunculan anotasi oleh sistem yang dilakukan berdasakan hasil *tokenizing* dan pengolahan token dengan rule yang telah dibangun. Berbeda dengan rule *name-nationality-gender*, rule email tidak melibatkan gazeteer apapun melainkan dengan mengandalkan logika algoritma yang membaca pola *token*. Performa dari rule ini diukur dari tingkat akurasi anotasi yang bernilai benar sesuai dengan label input data.

Gambar 6.4 Sampel Anotasi Input Email

Gambar 6.4 diatas merupakan sampel dari input data yang telah diberi anotasi email (*highlight* berwarna kuning). Dari 1000 input data yang diolah, seluruhnya dapat diberi anotasi email yang berarti tingkat akurasi dari rule email mencapai 100%.

6.4.4. Verifikasi Rule Phone

Verifikasi rule phone dilakukan dengan melihat kemunculan anotasi oleh sistem yang dilakukan berdasakan hasil *tokenizing* pengolahan oleh rule yang telah dibangun, dan beberapa format telepon membutuhkan penyocokkan dengan data gazeteer. Performa dari rule ini diukur dari tingkat akurasi anotasi yang bernilai benar sesuai dengan label input data.

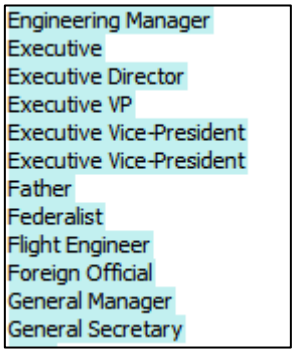


Gambar 6.5 Sampel Anotasi Input Phone

Gambar 6.5 diatas merupakan sampel dari input data yang telah diberi anotasi phone (*highlight* berwarna biru tua). Dari 1000 input data yang diolah, seluruhnya dapat diberi anotasi phone yang berarti tingkat akurasi dari rule phone mencapai 100%.

6.4.5. Verifikasi Rule Profesi

Verifikasi rule profesi dilakukan dengan melihat kemunculan anotasi oleh sistem yang dilakukan berdasakan hasil *tokenizing* pengolahan oleh rule yang telah dibangun, dan penyocokkan input data dengan data gazeteer. Performa dari rule ini diukur dari tingkat akurasi anotasi yang bernilai benar sesuai dengan label input data.

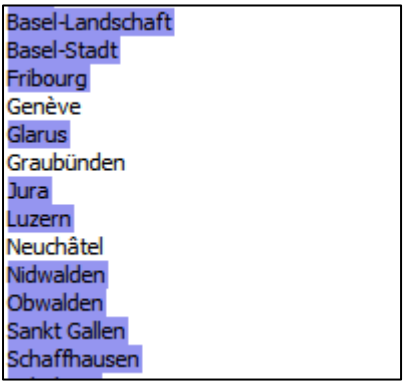


Gambar 6.6 Sampel Anotasi Input Profesi

Gambar 6.6 diatas merupakan sampel dari input data yang telah diberi anotasi profesi (*highlight* berwarna biru). Dari 1000 input data yang diolah, seluruhnya dapat diberi anotasi profesi yang berarti tingkat akurasi dari rule phone mencapai 100%.

6.4.6. Verifikasi Rule Address

Verifikasi rule address dilakukan dengan melihat kemunculan anotasi oleh sistem yang dilakukan berdasakan hasil *tokenizing* pengolahan oleh rule yang telah dibangun, dan penyocokkan input data dengan data gazeteer. Seluruh input data memiliki sekumpulan jenis address yang berbeda-beda sesuai dengan perancangan gazeteer jenis address. Performa dari rule ini diukur dari tingkat akurasi anotasi yang bernilai benar sesuai dengan label input data.



Gambar 6.7 Sampel Anotasi Input Address

Gambar 6.7 diatas merupakan sampel dari input data yang telah diberi anotasi address (*highlight* berwarna biru). Tingkat akurasi dari rule ini mencapai 82.9%, sebab dari 1000 input data hanya 829 yang dapat diberi anotasi. Hal ini dikarenakan JAPE tidak dapat mengolah karakter khusus walaupun data input terdapat pada data gazeteer. Berikut ini pada gambar 6.8 adalah contoh input data yang tidak dapat diolah dengan JAPE.

Genève
Graubünden
Gréivemaacher
Atlántico
Bolívar
Boyacá

Gambar 6.8 Sampel Input Address Tidak Bisa Diolah

6.5. Validasi Percobaan

Pada subbab ini mencakup penjelasan hasil percobaan sistem dengan menggunakan data fitur review dari layanan konten digital Google Play. Hasil percobaan menentukan apakah aplikasi dapat memberi bantuan dalam melakukan

pemetaan demografi pelanggan secara otomatis sesuai dengan tujuan penelitian. Validasi dilakukan dengan menggunakan muatan input data dari Google Play dan berdasarkan muatan data tersebut dilakukan pengamatan hasil anotasi yang dihasilkan oleh sistem. Dari hasil anotasi yang dihasilkan dilakukan rekap data untuk melihat informasi demografi pelanggan.

6.5.1. Muatan Input Data Percobaan

Dilakukan validasi sistem dengan objek penelitian aplikasi Youtube dengan banyak data sebesar 2491 data. Ketika dilakukan pemuatan input teks, terdapat data-data nama pelanggan yang kosong dimana akan membuat program tidak dapat mengintepretasi beberapa atribut demografi pelanggan. Maka data yang memiliki kondisi demikian tidak akan dimasukkan dalam pemrosesan berikutnya.

6.5.2. Kewarganegaraan dan Jenis Kelamin Pelanggan

Gambar 6.9 berikut merupakan contoh input data ulasan pelanggan yang terdapat anotasi nama. Anotasi tersebut diintrepretasikan sebagai nama wanita dengan jenis kewarganegaraan United States.

```
16772   1       Jessica Peter
        https://lh5.googleusercontent.com/-
De4ZpdpgeEY/AAAAAAAAAAI/AAAAAAAAAAA/AA6ZPT5FHpzwScIc-ZEpDwPwai2pu43106
https://play.google.com/store/apps/details?id=com.google.android.youtube&amp;reviewId=Z3A6QU9xcFRPSFAwMk9lamNoOG1FYjlfb0ZzOGM2bFhTTmVTTWxGSTAwMXBBN3dicGxud3Zvd3N2NnpjdWtDVU02MmlFdHR5MEdRVG9xS21IbHlxQX      3
                I enjoy youtube
```

Gambar 6.9 Cotoh Input Yang Diberi Anotasi

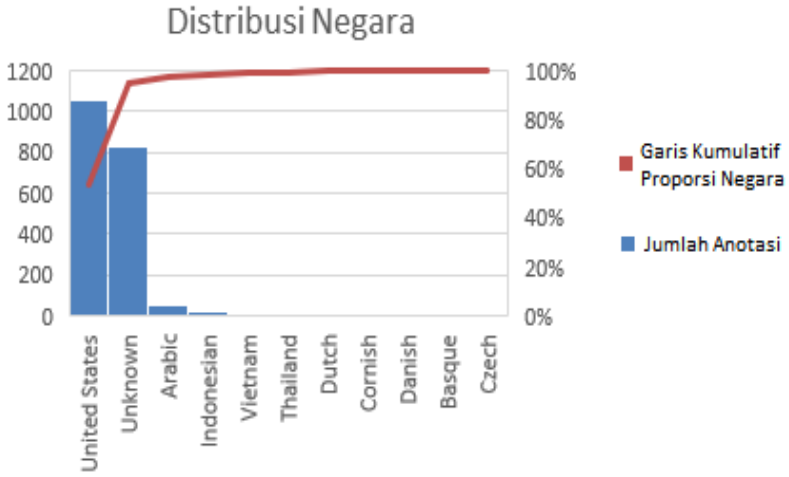
Tabel 6.8 merupakan rekap data statistik dari anotasi jenis kewarganegaraan dan jenis kelamin yang muncul setelah pengolahan input data review pelanggan aplikasi Youtube.

Tabel 6.8 Rekap Data Anotasi Kewarganegaraan dan Jenis Kelamin

Nationality	Gender	Total Per Jenis Kelamin	Total Keseluruhan
Arabic	Male	46	53
	Female	7	
Basque	Male	2	2
	Female	0	
Bulgarian	Male	0	0
	Female	0	
Catalan	Male	0	0
	Female	0	
Cornish	Male	3	3
	Female	0	
Croatian	Male	0	0
	Female	0	
Czech	Male	1	1
	Female	0	
Danish	Male	1	3
	Female	2	
Dutch	Male	4	4
	Female	0	
United States	Male	892	1053
	Female	161	
Finnish	Male	0	0

Nationality	Gender	Total Per Jenis Kelamin	Total Keseluruhan
	Female	0	
Indonesian	Male	16	19
	Female	3	
Thailand	Male	5	7
	Female	2	
Vietnam	Male	5	10
	Female	5	
Unknown	-	-	830
Total	-	-	1985
Tidak Terdeteksi			506

Berdasarkan tabel 6.8 diatas maka dapat dilihat proporsi kewarganegaraan pelanggan yang memberikan ulasan pada produk aplikasi Youtube sebagai berikut:

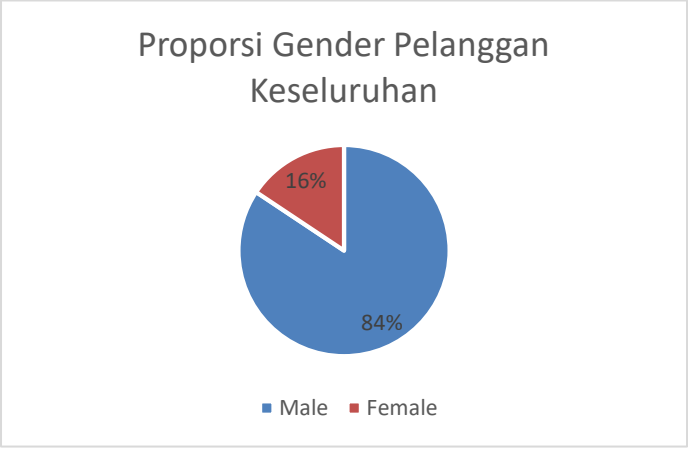


Gambar 6.10 Distribusi Kewarganegaraan Pelanggan

Dari visualisasi gambar 6.10 diatas menjelaskan distribusi anotasi kewarganegaraan pelanggan yang dirutkan berdasarkan jumlah besar ke kecil serta garis kumulatif (berwarna merah) yang menjelaskan persentase anotasi kewarganegaraan yang ditotalkan. Berikut ini merupakan presentase dari masing-masing kewarganegaraan pelanggan berdasarkan visualisasi diatas:

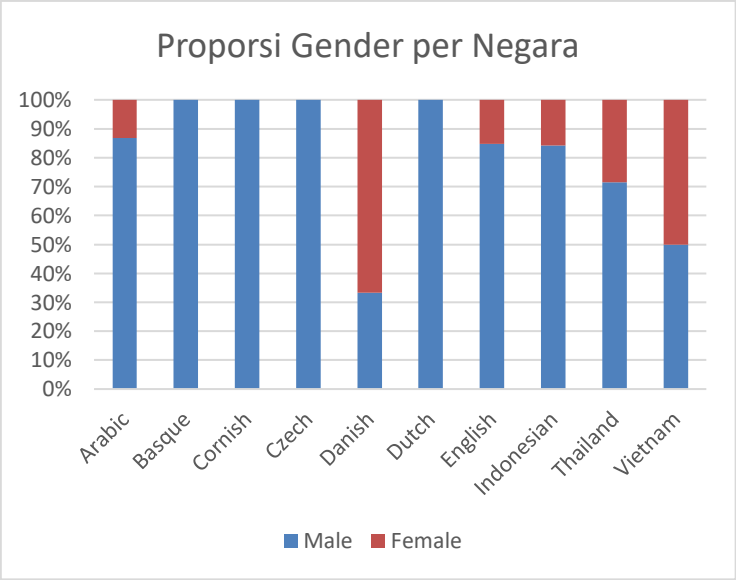
- United States 53%
- Unknown 41.8%
- Arabic 2.7%
- Indonesian 0.96%
- Vietnam 0.5%
- Thailand 0.35%
- Dutch 0.2%
- Cornish 0.15%
- Basque 0.1%
- Czech 0.05%

Terdapat data pelanggan dengan identifikasi warga negara yang belum diketahui (*Unknown*) sebesar 41.8%, hal ini terjadi dikarenakan sistem dapat membaca dan mendeteksi pola nama pada input data namun tidak dapat menemukan pada data gazeteer.Berdasarkan informasi ini dapat disimpulkan bahwa pengguna aplikasi Youtube yang memberi ulasan sebagian besar merupakan warga negara United States dengan persentase yang mencapai 53%.



Gambar 6.11 Proporsi Gender Pelanggan Secara Keseluruhan

Visualisasi pada gambar 6.11 diatas menjelaskan proporsi jenis kelamin pelanggan secara keseluruhan dari total 1155 input data yang teridentifikasi kewarganegaraannya. Dapat dilihat bahwa proporsi jenis kelamin pria jauh lebih besar daripada jenis kelamin wanita sehinga dapat disimpulkan pelanggan yang memberikan ulasan sebagian besar adalah pria.



Gambar 6.12 Proporsi Jenis Kelamin Pelanggan per Negara

Visualisasi gambar 6.12 diatas menjelaskan proporsi jenis kelamin dari setiap negara yang teridentifikasi oleh sistem. Didapatkan proporsi jenis kelamin dari setiap negara sebagai berikut:

- **Arabic:**
  Male: 86.8%, Female: 13.2%
- **Basque:**
  Male: 100%, Female: 0%
- **Cornish:**
  Male: 100%, Female: 0%
- **Czech**

Male: 0%, Female: 100%

- **Danish**
  Male: 33.3%, Female: 66.7%
- **Dutch**
  Male: 100%, Female: 0%
- **United States**
  Male: 85%, Female: 15%
- **Indonesian**
  Male: 84%, Female: 16%
- **Thailand**
  Male: 71%, Female: 29%
- **Vietnam**
  Male: 50%, Female: 50%

Berdasarkan data diatas, dapat diambil informasi bahwa setiap negara memiliki pola distribusi yang berbeda, sebagian besar didominasi oleh jenis kelamin pria, negara Denmark yang didominasi oleh wanita, dan Vietnam yang terdistribusi secara sama rata.

6.5.3. Asal dan Tempat Tinggal Pelanggan

Selanjutnya masuk ke pembahasan *address* yang terdeteksi oleh sistem. Tabel 6.9 berikut merupakan rekap data statistik dari anotasi *address* yang muncul setelah pengolahan input data ulasan pelanggan:

Tabel 6.9 Rekap Data Anotasi Address

Input yang dilabeli	Jumlah
US	4
Ra	3
Kannada	1
Paul	4
NY	3
Kanal	1
Rivera	1

Input yang dilabeli	Jumlah
America	1
Ig	1
England	1
Ba	2
Kara	1
Herrera	1
Monaghan	1
Jordan	1
Chin	1
Meta	1
Laguna	1
Mali	1
In Android	1
Mon	1
Vincent	1
Mo.	1
Kent	1
Total	35

Dari total input data sebanyak 2491 data, terdeteksi anotasi address sebanyak 35 anotasi. Seluruh dari input data yang diberi anotasi *address* bersifat bias. Berikut ini merupakan jenis dari bias *address* apa saja yang muncul beserta contoh data yang bias yang diberi warna kuning*:*

Tabel 6.10 Jenis Bias Anotasi Address

Jenis Bias	Kode Bias	Contoh Bias
Anotasi muncul berasal dari teks URL.	**B1**	https://play.google.com/store/apps/details?id=com.google.android.youtube&reviewId=Z3A6US

Jenis Bias	Kode Bias	Contoh Bias
Anotasi muncul berasal dari teks nama pelanggan.	**B2**	Jade Rivera
Anotasi berasal dari ekspresi komentar pelanggan yang tidak menggambarkan alamat	**B3**	Logan Paul amirite boyes? Ok, for real, the app is ok
Anotasi muncul karena terdapat token “in”.	**B4**	Previous updates I can find captions in English

Pada tabel 6.11 berikut menjelaskan input apa saja yang dilabeli beserta jenis bias yang terkait:

Tabel 6.11 Jumlah dan Kode Bias Address

Input yang dilabeli	Jumlah	Jumlah & Kode Bias
US	4	• 4 B1
Ra	3	• 3 B1
Kannada	1	• 1 B2
Paul	4	• 3 B4 • 1 B2
NY	3	• 3 B1
Kanal	1	• 1 B2
Rivera	1	• 1 B2
America	1	• 1 B2
Ig	1	• 1 B1
English	1	• 1 B4
Ba	2	• 2 B1
Kara	1	• 1 B2
Herrera	1	• 1 B2

Input yang dilabeli	Jumlah	Jumlah & Kode Bias
Monaghan	1	• 1 B2
Jordan	1	• 1 B2
Chin	1	• 1 B2
Meta	1	• 1 B2
Laguna	1	• 1 B2
Mali	1	• 1 B2
Android	1	• 1 B4
Mon	1	• 1 B2
Vincent	1	• 1 B2
Mo.	1	• 1 B2
Kent	1	• 1 B2
Total Bias	35	
Total Tidak Bias	0	

Dari total 35 input data yang diberi anotasi *address*, seluruhnya bersifat bias, sehingga data yang telah dilabeli anotasi tidak dapat dijadikan pedoman pengukuran asal atau tempat tinggal pelanggan.

6.5.4. Profesi Pelanggan

Selanjutnya masuk ke pembahasan profesi pelanggan yang terdeteksi oleh sistem. Dari total input data sebanyak 2491, terdeteksi sebanyak 20 anotasi profesi pelanggan. Berikut merupakan rekap data statistik dari anotasi profesi yang muncul setelah pengolahan input data ulasan pelanggan:

Tabel 6.12 Rekap Data Anotasi Profesi Pelanggan

Input yang dilabeli	Jumlah
King	4
Boss	1
Captain	1
Md	9
Editor	1
Minister	1

Input yang dilabeli	Jumlah
Commander	1
Carpenter	1
Jd	1
Total	20
Total	20

Dari total input data sebanyak 2491 data, terdeteksi anotasi profesi sebanyak 20 anotasi. Sebanyak 12 diantaranya anotasi bersifat bias, sedangkan sisanya sebanyak 7 anotasi dapat dikatakan rancu, apakah bersifat bias atau benar menggambarkan profesi. Berikut ini merupakan jenis dari bias profesi apa saja yang muncul beserta contoh data bias yang diberi warna kuning:

Tabel 6.13 Jenis Bias Anotasi Profesi

Jenis Bias	Kode Bias	Contoh Bias
Anotasi muncul berasal dari teks URL.	**B1**	https://lh5.googleusercontent.com/-W1vXGer4qyE/AAAAAAAAAAI/AAAAAAAAAAB4/5Jd4dg1X-7w/
Anotasi muncul berasal dari teks nama pelanggan.	**B2**	Meta Carpenter
Anotasi berasal dari ekspresi komentar pelanggan yang tidak menggambarkan profesi	**B3**	I think this is best like a boss

Pada tabel 6.14 berikut menjelaskan input apa saja yang dilabeli beserta jenis bias yang terjadi:

Tabel 6.14 Jumlah dan Kode Bias Profesi

Input yang dilabeli	Jumlah	Jumlah & Kode Bias
King	4	• 4 B2
Boss	1	• 1 B2
Captain	1	• 1 B2
Md	2	• 1 B1 • 1 B3
Editor	1	• 1 B3
Minister	1	• 1 B3
Commander	1	• 1 B3
Carpenter	1	• 1 B2
Jd	1	• 1 B1
Total Bias	12	
Total Tidak Bias	8	

Terdapat total 7 anotasi yang rancu, dikatakan demikian sebab secara makna dari input data yang telah diberi anotasi harus dipertanyakan apakah yang dimaksud menggambarkan profesi dari pelanggan terkait. Berikut adalah input data yang diberi anotasi profesi beserta token yang dapat mendukung untuk melihat apakah anotasi bersifat bias atau tidak:

Tabel 6.15 Anotasi Profesi Yang Tidak Bias

Md. Rohul Amin
Md Nazmul
Md Nur
Md. Abdur Rahman
Md Bokul
Md Sakib Al Hasan
Md Iqbal

Anotasi diatas dapat dikatakan benar apabila makna dari “Md” adalah gelar “*Doctor of Medicine Profession* (Md)”. Kemudian dapat dikatakan bias apabila Md merupakan bagian dari nama pelanggan.

6.5.5. Nomor Telepon Pelanggan

Selanjutnya masuk ke pembahasan nomor telepon pelanggan yang terdeteksi oleh sistem. Dari total input data sebanyak 2491, terdeteksi sebanyak 10 anotasi nomor telepon pelanggan dan seluruhnya bersifat bias. Berikut merupakan rekap data statistik dari anotasi nomor telepon yang muncul setelah pengolahan data ulasan pelanggan beserta alasan mengapa data bersifat bias:

Tabel 6.16 Anotasi Phone dan Alasan Bias

Input yang dilabeli	Alasan Bias
Error code 501 1690	Bukan ekspreksi pengungkapan nomor telepon.
Its not installing on my micromax q335 16934	Bukan ekspreksi pengungkapan nomor telepon.
18337	Merupakan nomor baris data.
Bad version hank mobile for honor 7x 18541	Bukan ekspreksi pengungkapan nomor telepon.
https://lh5.googleusercontent.com/-uAaB0in8d-0/AAAAAAAAAAI/AAAAAAAAAAA/AA6ZPT6IOGDg25pXDFRoJX0xyx032 43105	Bagian dari URL.
https://lh5.googleusercontent.com/-uAaB0in8d-0/AAAAAAAAAAI/AAAAAAAA	Bagian dari URL.

Input yang dilabeli	Alasan Bias
AAA/AA6ZPT6IOGDg25pXDFRoJX0xyx032 43105	
https://lh3.googleusercontent.com/-NwiOE6lpagU/AAAAAAAAAAI/AAAAAAAAAAA/AA6ZPT7lr5Nw411Hvy-CJCukr719x 43105	Bagian dari URL.
https://lh3.googleusercontent.com/-NwiOE6lpagU/AAAAAAAAAAI/AAAAAAAAAAA/AA6ZPT7lr5Nw411Hvy-CJCukr719x43105	Bagian dari URL.
https://lh6.googleusercontent.com/-ACdtmF8QxCo/AAAAAAAAAAI/AAAAAAAAAAA/AA6ZPT5UIFWiRSf6ujMdlBwSC0025 43105	Bagian dari URL.
19061	Merupakan nomor baris data.
Cant download videos from youtube anymore and after i press view more replies, i cant close the reply section by clicking the x.	Bukan ekspreksi pengungkapan nomor telepon.

6.5.6. Ringkasan Demografi Pelanggan

Dengan menggunakan sistem yang dibangun maka didapatkan informasi bahwa sebagian besar pelanggan aplikasi Youtube merupakan warga negara United States dengan nilai proporsi mencapai 53%. Selanjutnya terdapat 41.8% dari total input data yang belum diketahui jenis kewarganegaraannya, hal ini dikarenakan data gazeteer nama belum sepenuhnya memadai untuk mendeteksi seluruh nama pelanggan. Untuk meningkatkan performa identifikasi kewarganegaraan, maka gazeteer nama perlu diperkaya lagi. Selanjutnya untuk jenis kelamin pelanggan, sebagian besar merupakan pria dengan nilai proporsi sebesar 84% dari total seluruh anotasi nama yang terdeteksi, tidak termasuk yang kewarganegaraannya tidak diketahui. Sisanya sebesar 16% merupakan wanita. Kemudian

untuk hasil identifikasi terkait dengan atribut *address, phone* hampir seluruhnya bersifat bias, dan terdapat 8 anotasi profesi yang tidak bersifat bias dari total identifikasi sebanyak 20 anotasi.

BAB VII
KESIMPULAN DAN SARAN

Pada bab ini dibahas mengenai kesimpulan dari semua proses yang telah dilakukan dan saran yang dapat diberikan untuk pengembangan yang lebih baik.

7.1. Kesimpulan

Berdasarkan pengerjaan tugas akhir dengan judul "Prototipe Ekstraksi Profil Demografi Pelanggan Menggunakan Metode Template Filling Untuk Personalisasi Pencarian Produk" yang telah dilakukan dapat disimpulkan beberapa hal sebagai berikut:

1. Template dibangun berdasarkan pemetaan penelitian berjudul "*Ontology-Based User Modeling for E-Commerce System*" yang ditulis oleh Weilong Liu, Fang Jin, Xin Zhang [20] dengan ketersediaan data pada Google Play. Adapun atribut template demografi pelanggan yang dihasilkan dari proses pemetaan mencakup *Nationality, Gender, Profession Email, Address,* dan *Phone.* Template ini tidak sepenuhnya cocok dengan lingkungan implementasi (Google Play). Adapun atribut template yang tidak cocok dengan lingkungan implementasi antara lain:
   a. Email: Dari sekian banyak input data yang diolah, tidak terdapat anotasi email sama sekali. Sehingga atribut ini dinilai tidak cocok dengan lingkungan implementasi.
   b. Phone Number: Hampir pasti anotasi yang dihasilkan bersifat bias. Ini dikarenakan konten input data mencakup URL dan pola-pola angka yang menyerupai nomor telepon.
2. Metode template filling dengan pendekatan *finite-state cascade* pada penelitian ini memiliki 3 komponen utama. Yang pertama adalah *rule*, berperan sebagai sistem berbasis aturan untuk membaca pola input dan mengintepretasi anotasi sesuai pola yang terdeteksi. *Rule* mencakup pemrograman JAPE yang telah dibangun dan didukung

dengan aplikasi ANNIE yang telah disediakan oleh GATE Developer (*sentence splitter*, POS, *gazeteer loader*, dan *tokeniser*). Yang kedua adalah *gazeteer* yang berperan sebagai kamus data sistem identifikasi, cara kerja kamus data tergantung dengan *rule* yang dibangun apakah *gazeteer* dideklarasi atau tidak. Dan yang terakhir adalah input teks yang akan diidentifikasi, pada penelitian input teks adalah data ulasan pelanggan Google Play.

3. Identifikasi profil demografi pelanggan dengan menggunakan template menghasilkan hasil yang berbeda-beda. Untuk atribut kewarganegaraan, dan jenis kelamin dapat berjalan dengan cukup baik walaupun terdapat kekurangan yaitu gazeteer yang masih bisa diperkaya lebih banyak lagi agar identifikasi demografi pada input teks dapat lebih banyak lagi. Untuk atribut *address* hampir seluruh input data data yang diberi anotasi bersifat bias. Untuk atribut profesi terdapat anotasi yang dapat dijadikan acuan yang menggambarkan profesi pelanggan walaupun masih terdapat kekurangan. Untuk atribut *phone number*, seluruh anotasi yang dihasilkan bersifat bias. Sedangkan untuk atribut email tidak muncul sedikitpun dari sekian banyak input data. Dengan menggunakan sistem yang dibangun, maka didapatkan informasi bahwa:
   - Dari total input data yang diolah, sebagian besar pelanggan memiliki kewarganegaraan United States dengan nilai proporsi 53%.
   - Terdapat 41.8% dari total data yang belum diketahui jenis kewarganegaraannya. Untuk meningkatkan performa identifikasi kewarganegaraam, maka gazeteer kewarganegaraan perlu diperkaya lagi.
   - Dari keseluruhan input data nama yang dapat diberi anotasi kewarganegaraan, sebagian besar pelanggan yang memberi ulasan merupakan pria dengan nilai 84%.
   - Masing-masing kewarganegaraan yang teridentifikasi memiliki proporsi nilai yang berbeda, namun sebagian besar dari setiap pelanggan merupakan pelanggan

berjenis kelamin pria. Hanya negara Denmark saja yang jumlah wanita lebih banyak dari pria.

- Address pelanggan yang teridentifikasi seluruhnya bersifat bias. Sehingga perlu sampel input data yang lebih untuk melihat apakah terdapat ekspresi ungkapan tempat tinggal pelanggan baik secara langsung, atau tidak langsung (contoh: there are some contents that aren't available in **India**).
- Profesi yang diungkapkan oleh pelanggan tidaklah banyak jika dibandingkan dari total input data. Terdapat 8 profesi yang teridentifikasi dan seluruhnya sama persis (Md.).
- Dari keseluruhan input data, teridentifikasi sebanyak 10 data telepon pelanggan. Namun seluruhnya bersifat bias.

7.2. Saran

Saran penulis untuk penelitian dan pengembangan selanjutnya adalah sebagai berikut:

1. Produk pada penelitian ini berupa sekumpulan *rules* dapat dikembangkan lebih lanjut menjadi aplikasi tersendiri tanpa pengoperasian yang tergantung dengan GATE Developer dengan menggunakan API yang disediakan oleh GATE.
2. Pengambilan data dapat diperbarui lagi dengan menghubungkan langsung secara real time dari sumber data dengan aplikasi yang telah dikembangkan dengan API GATE.
3. Dapat dikembangkan sistem visualisasi berupa dashboard dengan menggunakan hasil anotasi profil demografi pelanggan yang teridentifikasi oleh sistem.
4. *Data dictionary* atau *gazeteer* dapat diperkaya lagi sehingga performa aplikasi dapat lebih akurat dalam mengintepretasi hasil identifikasi.

Halaman ini sengaja dikosongkan

DAFTAR PUSTAKA

[1] T. Wallace, "The Driving Factors Behind Modern Consumer Behavior in the New Omni-Channel World [Infographic]," *BIGCOMMERCE*. [Online]. Available: https://www.bigcommerce.com/blog/consumer-behavior-infographic/. [Accessed: 23-Feb-2017].

[2] T. P. Patrick Gibbons, Jeff Marr, Sonya McAllister, Leslie Pagel, "The Future of B-to-B Customer Experience 2020," 2014.

[3] K. Harrison, "Audience segmentation is important for better communication," *Cutting Edge PR*. [Online]. Available: http://www.cuttingedgepr.com/articles/audience-segmentation-better-communication.asp. [Accessed: 25-Feb-2017].

[4] S. Yang, "An Ontology-Supported User Modeling Technique with Query Templates for Interface Agents," *Interface*, pp. 556–561, 2007.

[5] K. K. Kotler, Philip, *Marketing Management 13th edition*, 13th ed. New Jersey: Pearson Education, 2009.

[6] W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, *Encyclopedia of Systems Biology*. 2013.

[7] N. Huynh and Q. Ho, "A Combined Approach for Disease/Disorder Template Filling," *2015 Seventh Int. Conf. Knowl. Syst. Eng.*, pp. 328–331, 2015.

[8] S. K. Endarnoto, S. Pradipta, A. S. Nugroho, and J. Purnama, "Traffic condition information extraction & visualization from social media twitter for android mobile application," *Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011*, no. July, pp. 0–3, 2011.

[9] R. E. P. Efraim Turban, R. Kelly Rainer, *Introduction to*

Information Technology, 3rd ed. Wiley, 2005.

[10] G. Venugopal, "A Review of Popular Applications on Google Play – Do They Cater to Visually Impaired Users ?," vol. 8, no. February 2014, pp. 221–239, 2015.

[11] E. Chu, "Android Market update: support for priced applications," 2013. [Online]. Available: https://android-developers.googleblog.com/2009/02/android-market-update-support-for.html.

[12] "The Science of Population," *demographicpartitions.org*, 2014. [Online]. Available: https://web.archive.org/web/20150814023915/http://demographicpartitions.org/science-population-determines-population-change/.

[13] G. G. Chowdhury, "Natural language processing," *Annu. Rev. Inf. Sci. Technol.*, vol. 37, no. 1, pp. 51–89, 2005.

[14] J. Pittsburgh, "Machine Learning for Information Extraction in Informal Domains," pp. 169–202, 2000.

[15] K. A. and R. C. M. Sasikumar, S. Ramani, S. M. Raman, "A Practical Introduction to Rule Based Expert Systems," no. November, 2007.

[16] B. M. N. and A. D. John D. Kelleher, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. London, England: MIT Press.

[17] N. Y. Wirawan, "Rancang Bangun Ekstraksi Topik Fitur Produk dari Ulasan Pengguna Online dengan Latent Dirichlet Allocation," 2017.

[18] M. D. Campbell, "Behind the Name: the Etymology and History of First Names.," *Behind the Name: the*

Etymology and History of First Names., 2013. [Online]. Available: http://www.behindthename.com/.

[19] United Nations Economic Commission for Europe (UNECE), 2017. [Online]. Available: http://www.unece.org/cefact/codesfortrade/codes_index.html

[20] W. Liu and F. Jin, X. Zhang " Ontology-Based User Modeling for E-Commerce System "

Halaman ini sengaja dikosongkan

BIODATA PENULIS

Penulis lahir di Ujung Pandang pada tanggal 1 September 1994. Merupakan anak kedua dari 2 bersaudara. Penulis telah menempuh beberapa pendidikan formal, yaitu; SD Bina Insani, SD Madania ISWCS, SMP Madania ISWCS, dan SMA Madania ISWCS.

Pada tahun 2013 pasca kelulusan SMA, penulis melanjutkan pendidikan dengan jalur mandiri di Jurusan Sistem Informasi Institut Teknologi Sepuluh Nopember (ITS) Surabaya dan terdaftar sebagai mahasiswa dengan NRP 5213100180.

Penulis memiliki ketertarikan pada bidang pengolahan data, oleh karena itu penulis memutuskan untuk mengambil minat Sistem Enterprise dengan topik penelitian berkaitan dengan pengolahan data. Untuk kepentingan penelitian selanjutnya, penulis dapat dihubungi melalui email di adisatria21@gmail.com

Halaman ini sengaja dikosongkan

LAMPIRAN A

Lampiran A mencakup file JAPE yang berisikan rule:

- File JAPE rule address tersedia dalam bentuk softcopy dengan nama file addressphy.jape.
- File JAPE rule email tersedia dalam bentuk softcopy dengan nama file email.jape.
- File JAPE rule profesi tersedia dalam bentuk softcopy dengan nama file job.jape.
- File JAPE rule name-nationality-gazeteer tersedia dalam bentuk softcopy dengan nama file name.jape.
- File JAPE rule phone tersedia dalam bentuk softcopy dengan nama file phone.jape.

Halaman ini sengaja dikosongkan

LAMPIRAN B

Lampiran B mencakup file Gazeteer yang digunakan pada penelitian:

- File Gazeteer nama Arab tersedia dalam bentuk softcopy dengan nama ArabicFNf.lst (Arabic Firstname Female), ArabicFNm.lst (Arabic Firstname Male), ArabicSNf.lst (Arabic Surname Female), ArabicSNm.lst (Arabic Surname Male).
- File Gazeteer nama Basque tersedia dalam bentuk softcopy dengan nama BasqueFNf.lst (Basque Firstname Female), BasqueFNm.lst (Basque Firstname Male), BasqueSNf.lst (Basque Surname Female), BasqueSNm.lst (Basque Surname Male).
- File Gazeteer nama Bulgaria tersedia dalam bentuk softcopy dengan nama BulgarianFNf.lst (BulgarianFirstname Female), BulgarianFNm.lst (BulgarianFirstname Male), BulgarianSNf.lst (BulgarianSurname Female), BulgarianSNm.lst (BulgarianSurname Male).
- File Gazeteer nama Catalan tersedia dalam bentuk softcopy dengan nama Catalan FNf.lst (Catalan Firstname Female), Catalan FNm.lst (Catalan Firstname Male), Catalan SNf.lst (Catalan Surname Female), CatalanSNm.lst (Catalan Surname Male).
- File Gazeteer nama Cornish tersedia dalam bentuk softcopy dengan nama Cornish FNf.lst (Cornish Firstname Female), Cornish FNm.lst (Cornish Firstname Male), Cornish SNf.lst (Cornish Surname Female), CornishSNm.lst (Cornish Surname Male).
- File Gazeteer nama Croatia tersedia dalam bentuk softcopy dengan nama CroatianFNf.lst (Croatian Firstname Female), CroatianFNm.lst (Croatian Firstname Male), CroatianSNf.lst (CroatianSurname Female), CroatianSNm.lst (Croatian Surname Male).
- File Gazeteer nama Czech tersedia dalam bentuk softcopy dengan nama CzechFNf.lst (Czech Firstname Female), CzechFNm.lst (Czech Firstname Male), CzechSNf.lst (Czech Surname Female), CzechSNm.lst (Czech Surname Male).
- File Gazeteer nama Denmark tersedia dalam bentuk softcopy dengan nama DanishFNf.lst (Danish Firstname Female), DanishFNm.lst (Danish Firstname Male), DanishSNf.lst

(DanishSurname Female), DanishSNm.lst (Danish Surname Male).

- File Gazeteer nama Dutch tersedia dalam bentuk softcopy dengan nama DutchFNf.lst (Dutch Firstname Female), DutchFNm.lst (Dutch Firstname Male), DutchSNf.lst (Dutch Surname Female), DutchSNm.lst (Dutch Surname Male).
- File Gazeteer nama Finland tersedia dalam bentuk softcopy dengan nama FinlandFNf.lst (Finland Firstname Female), FinlandFNm.lst (Finland Firstname Male), FinlandSNf.lst (Finland Surname Female), FinlandSNm.lst (Finland Surname Male).
- File Gazeteer nama Indonesia tersedia dalam bentuk softcopy dengan nama IndonesiaFNf.lst (Indonesia Firstname Female), IndonesiaFNm.lst (Indonesia Firstname Male), IndonesiaSNf.lst (Indonesia Surname Female), IndonesiaSNm.lst (Indonesia Surname Male).
- File Gazeteer nama Thailand tersedia dalam bentuk softcopy dengan nama Thailand FNf.lst (Thailand Firstname Female), ThailandFNm.lst (Thailand Firstname Male), ThailandSNf.lst (Thailand Surname Female), ThailandSNm.lst (Thailand Surname Male).
- File Gazeteer nama United States tersedia dalam bentuk softcopy dengan nama EnglishFNf.lst (United States Firstname Female), EnglishFNm.lst (United States Firstname Male), EnglishSNf.lst (United States Surname Female), EnglishSNm.lst (United States Surname Male).
- File Gazeteer nama Vietnam tersedia dalam bentuk softcopy dengan nama VietnamFNf.lst (Vietnam Firstname Female), VietnamFNm.lst (Vietnam Firstname Male), VietnamSNf.lst (Vietnam Surname Female), VietnamSNm.lst (Vietnam Surname Male).
- File Gazeteer profesi tersedia dalam bentuk softcopy dengan nama file profesi.lst.
- File Gazeteer Wilayah tersedia dalam bentuk softcopy dengan nama canton_address.lst, capitalcity_address.lst, councilarea_address.lst, countries_address.lst, department_address.lst, district_address.lst, municipality_address.lst, prefecture_address.lst, province_address.lst, region_address.lst, state_address.lst.

LAMPIRAN C

Lampiran C mencakup input data Google Play yang digunakan pada proses Validasi dan file yang digunakan dalam proses verifikasi. File input data Google Play tersedia dalam bentuk softcopy dengan nama input_youtube.csv, dan input_youtube2.csv. Untuk verifikasi disimpan secara terpisah dengan penjelasan seperti berikut:

- File verifikasi nama Arab tersedia dalam bentuk softcopy dengan nama file arab_female.txt dan arab_male.txt.
- File verifikasi nama Basque tersedia dalam bentuk softcopy dengan nama file Basque_female.txt dan Basque_male.txt.
- File verifikasi nama Bulgaria tersedia dalam bentuk softcopy dengan nama file bulgarian_female.txt dan bulgarian_male.txt.
- File verifikasi nama Catalan tersedia dalam bentuk softcopy dengan nama file catalan_female.txt dan catalan_male.txt.
- File verifikasi nama Cornish tersedia dalam bentuk softcopy dengan nama file cornish_female.txt dan cornish_male.txt.
- File verifikasi nama Croatia tersedia dalam bentuk softcopy dengan nama file croatian_female.txt dan croatian_male.txt.
- File verifikasi nama Czech tersedia dalam bentuk softcopy dengan nama file czech_female.txt dan czech_male.txt
- File verifikasi nama Denmark tersedia dalam bentuk softcopy dengan nama file danish_female.txt dan danish_male.txt.
- File verifikasi nama Dutch tersedia dalam bentuk softcopy dengan nama file dutch_female.txt dan dutch_male.txt.
- File verifikasi nama Finland tersedia dalam bentuk softcopy dengan nama file finnish_female.txt dan finnish_male.txt.
- File verifikasi nama Indonesia tersedia dalam bentuk softcopy dengan nama file indonesian_female.txt dan indonesian_male.txt.
- File verifikasi nama Thailand tersedia dalam bentuk softcopy dengan nama file thailand_female.txt dan thailand_male.txt.
- File verifikasi nama United States tersedia dalam bentuk softcopy dengan nama file unitedstates_female.txt dan unitedstates_male.txt.
- File verifikasi nama Vietnam tersedia dalam bentuk softcopy dengan nama file vietnam_female.txt dan vietnam_male.txt.
- File verifikasi address tersedia dalam bentuk softcopy dengan nama file address.txt.

- File verifikasi email tersedia dalam bentuk softcopy dengan nama file email.txt.
- File verifikasi phone tersedia dalam bentuk softcopy dengan nama file phone.txt.
- File verifikasi profession tersedia dalam bentuk softcopy dengan nama file profession.txt.