**BAB 2**

**TINJAUAN PUSTAKA**

* 1. **Landasan Teori**
     1. **Algoritma**

Menurut Skiena (2008), Algoritma adalah sebuah prosedur untuk menyelesaikan suatu pekerjaan atau masalah. Untuk menjadi sesuatu yang menarik, algoritma harus dapat menyelesaikan suatu masalah yang umum dan di spesifikasi dengan baik.

Algoritma berisi langkah-langkah penyelesaian suatu masalah. Langkah-langkah tersebut dapat berupa runtunan aksi, pemilihan aksi, dan pengulangan aksi. Ketiga jenis langkah tersebut membentuk konstruksi suatu algoritma. Jadi, sebuah algoritma dapat dibangun dari tiga buah struktur dasar, yaitu :

1. Runtunan (Sequence)

Sebuah runtunan terdiri dari satu atau lebih instruksi. Tiap instruksidikerjakan secara berurutan sesuai dengan urutan penulisannya, yakni sebuah instruksi dilaksanakan setelah instruksi sebelumnya selesai dikerjakan.

1. Pemilihan (Selection)

Adakalanya sebuah instruksi dikerjakan jika kondisi tertentu dipenuhi. Tiap–tiap instruksi akan diseleksi oleh kondisi, apabila instruksi memenuhi kondisi yang diminta, maka instruksi akan dijalankan.

1. Pengulangan (Repetition)

Salah satu kelebihan komputer adalah kemampuannya untuk mengerjakan pekerjaan yang sama berulang kali tanpa mengenal lelah. Kita tidak perlu menulis instruksi yang sama berulang kali, tetapi cukup melakukan pengulangan dengan instruksi yang tersedia.

* + 1. **Artificial Intelligence**

Menurut Nilsson (2009), artificial intelligence adalah semua yang ditujukan untuk membuat sebuah mesin yang cerdas dimana kecerdasan yang dimaksud adalah yang memungkinkan suatu entitas dapat berfungsi secara tepat dengan pandangan terdepan di dalam lingkungannya.

Perbandingan Kecerdasan Buatan dan Kecerdasan Manusia menurut Kaplan (Turban, McLean, dan Wetherbe, 1999), AI mempunyai beberapa kelebihan dibandingkan dengan kecerdasan alami/manusia. Adapun kelebihan AI sebagai berikut:

* AI bersifat Permanen, sepanjang sistem dan program masih terpelihara, sedangkan kecerdasan alami seseorang tidak dapat disimpan.
* AI menawarkan kemudahan untuk digandakan dan disebarkan.
* Pengetahuan dalam AI dapat lebih murah dibanding kecerdasan alami/manusia, biaya membeli jasa dengan computer lebih murah dibanding membiayai manusia dengan tugas yang sama.
* AI bersifat konsisten dan teliti, sedangkan manusia tidak konsisten.
* AI dapat didokumentasikan.
  + 1. **Natural Language Processing (NLP)**

Natural Language adalah bahasa-bahasa yang biasa digunakan oleh manusia. Untuk dapat memproses bahasa-bahasa tersebut ke dalam komputer maka terbentuklah sebuah ilmu yang dinamakan Natural Language Processing. Menurut Abhimanyu Chopra (2013), Natural Language Processing adalah bagian dari ilmu artificial intelligence dan linguistic, dibentuk untuk membuat komputer dapat mengerti pernyataan atau kata-kata yang ditulis dalam bahasa manusia.

NLP terdiri dari beberapa bagian utama yaitu:

* Parser, suatu sistem untuk mengambil kalimat input, dan menguraikannya kata per kata serta untuk menentukan jenis kata apa saja yang dapat mengikuti kata tersebut. (Menguraikan kalimat ke dalam beberapa bagian gramatikal seperti subjek, objek, kata kerja, kata benda, kata sifat dan sebagainya).
* Sistem Representasi Pengetahuan, suatu sistem untuk menganalisa hasil dari parser dan menentukan maknanya.
* Output Translator, suatu hasil terjemahan yang merepresentasikan sistem pengetahuan serta melakukan langkah-langkah yang berupa jawaban atas bahasa alami. Output translator merupakan output khusus yang sesuai dengan program komputer lainnya.

Pustejovsky dan Stubbs (2012) menjelaskan bahwa ada beberapa area utama penelitian pada field NLP, diantaranya:

1. Question Answering Systems (QAS)

Kemampuan komputer untuk menjawab pertanyaan yang diberikan oleh user. Daripada memasukkan keyword ke dalam browser pencarian, dengan QAS, user bisa langsung bertanya dalam bahasa natural yang digunakannya, baik itu Inggris, Mandarin, ataupun Indonesia.

1. Summarization

Pembuatan ringkasan dari sekumpulan konten dokumen atau email. Dengan menggunakan aplikasi ini, user bisa dibantu untuk mengkonversikan dokumen teks yang besar ke dalam bentuk slide presentasi

1. Machine Translation

Produk yang dihasilkan adalah aplikasi yang dapat memahami bahasa manusia dan menerjemahkannya ke dalam bahasa lain. Salah satu contoh yang dapat kita lihat adalah Google Translate yang telah menjadi alat penerjemah di dalam kehidupan sehari-hari jutaan orang di berbagai belahan dunia.

1. Speech Recognition

Field ini merupakan cabang ilmu NLP yang cukup sulit. Namun saat ini telah banyak dikembangkan speech recognition untuk digunakan di smartphone atau komputer dalam mengenali bahasa yang diucapkan, yang biasanya berupa pernyataan dan perintah.

1. Document Classification

Field ini dapat dikatakan sebagai area penelitian NLP yang paling sukses dan telah memberi kemudahan bagi kita untuk mengklasifikasikan dokumen berdasarkan kategori yang ada. Document classification sangat berguna digunakan pada aplikasi seperti email untuk menyaring spam, klasifikasi kategori artikel atau berita, dan ulasan film.

* + 1. **Machine Learning**

Machine Learning adalah cabang dari ilmu kecerdasan buatan (artificial intelligence) yang berfokus pada pembangunan dan studi sebuah sistem agar mampu belajar dari data-data yang diperolehnya. Menurut Arthur Samuel, Machine Learning adalah bidang studi yang memberikan kemampuan program komputer untuk belajar tanpa secara eksplisit di program.

Machine learning sebagai sebuah solusi dapat digunakan jika sebuah kasus masalah memenuhi tiga hal sebagai berikut

1. A Pattern exist (terdapat sebuah pola). Jika data yang ada bersifat acak. Maka machine learning tidak akan dapat menyelesaikan permasalahan tersebut.
2. Persoalan tersebut tidak dapat diselesaikan secara matematis. Machine learning menggunakan teori ilmu statistic (probabilitas dan pendekatan). Sehingga ketika sebuah peristiwa tidak dapat diselesaikan dengan pendekatan perhitungan matematik biasa, maka Machine learning-lah solusinya.
3. Terdapat data. Sebagaimana namanya machine learning adalah sebuah proses dimana sebuah aplikasi sanggup belajar dari data yang diberikan. Jika tidak terdapat data, maka machine learning tidak akan bisa menyelesaikan masalah tersebut.

Teknik dari machine learning dibagia menjadi tiga kategori yaitu supervised learning, unsupervised learning dan reinforcement learning (Russell & Norvig, 2010).

Supervised learning merupakan suatu teknik yang diggunakan untuk mendapatkan beberapa metode yang akan digunakan sebagai classifier untuk memprediksi input yang diberikan. Beberapa contoh dari algoritma supervised learning adalah Naïve Bayes, SVM (Support Vector Machine), dan Neural Networks.

Unsupervised learning merupakan teknik yang sama dengan supervised learning. Sama halnya dengan supervised, unsupervised learning juga berfungsi sebagai metode classifier untuk memprediksi input yang diberikan. Tetapi perbedaannya terletak pada metode evaluasi dimana pada unsupervised learning tidak dilakukan perbandingan dengan data manual, melainkan model yang merupakan hasil prediksi dengan masing-masing data.

Reinforcement learning merupakan metode pembelajaran terhadap apa yang akan dilakukan atau bagaimana memetakan situasi terhadap aksi yang akan dilakukan oleh sistem untuk mendapatkan reward yang maksimal. Metode ini berbeda dengan 2 metode lainnya dimana metode ini tidak membutuhkan skenario dari aksi yang akan dilakukan, tetapi lebih pada menentukan aksi yang dapat memberikan reward yang maksimal.

* + 1. **Support Vector Machine**

Menurut Santoso (2007) Support vector machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. SVM berada dalam satu kelas dengan Artificial Neural Network (ANN) dalam hal fungsi dan kondisi permasalahan yang bisa diselesaikan. Keduanya masuk dalam kelas supervised learning.

Kelebihan  Dalam memilih solusi untuk menyelesaikan suatu masalah, kelebihan dan kelemahan masing-masing metode harus diperhatikan. Selanjutnya metode yang tepat dipilih dengan memperhatikan karakteristik data yang diolah. Dalam hal SVM, walaupun berbagai studi telah menunjukkan kelebihan metode SVM dibandingkan metode kon vensional lain, SVM juga memiliki berbagai kelemahan. Kelebihan SVM antara lain sbb.

1.   Generalisasi

Generalisasi didefinisikan sebagai kemampuan suatu metode (SVM, neural network, dsb.) untuk mengklasifikasikan suatu pattern, yang tidak termasuk data yang dipakai dalam fase pembelajaran metode itu. Vapnik menjelaskan bahwa generalization error dipengaruhi oleh dua faktor: error terhadap training set, dan satu faktor lagi yang dipengaruhi oleh dimensi VC (Vapnik-Chervokinensis). Strategi pembelajaran pada neural network dan umumnya metode learning machine difokuskan pada usaha untuk meminimimalkan error pada training-set. Strategi ini disebut Empirical Risk Minimization (ERM). Adapun SVM selain meminimalkan error pada training-set, juga meminimalkan faktor kedua. Strategi ini disebut Structural Risk Minimization (SRM), dan dalam SVM diwujudkan dengan memilih hyperplane dengan margin terbesar. Berbagai studi empiris menunjukkan bahwa pendekatan SRM pada SVM memberikan error generalisasi yang lebih kecil daripada yang diperoleh dari strategi ERM pada neural network maupun metode yang lain.

2.  Curse of dimensionality

Curse of dimensionality didefinisikan sebagai masalah yang dihadapi suatu metode pattern recognition dalam mengestimasikan parameter (misalnya jumlah hidden neuron pada neural network, stopping criteria dalam proses pembelajaran dsb.) dikarenakan jumlah sampel data yang relatif sedikit dibandingkan dimensional  ruang vektor data te rsebut. Semakin tinggi dimensi dari ruang vektor informasi yang diolah, membawa konsekuensi dibutuhkannya jumlah data dalam proses pembelajaran. Pada kenyataannya seringkali terjadi, data yang diolah berjumlah terbatas, dan untuk mengumpulkan data yang lebih banyak tidak mungkin dilakukan karena kendala biaya dan kesulitan teknis. Dalam kondisi tersebut, jika metode itu “terpaksa” harus bekerja pada data yang berjumlah relatif sedikit dibandingkan dimensinya, akan membuat proses estimasi parameter metode menjadi sangat sulit.

3.   Landasan teori

Sebagai metode yang berbasis statistik,  SVM memiliki landasan teori yang dapat dianalisa dengan jelas, dan tidak bersifat black box.

4.   Feasibility

SVM dapat diimplementasikan relatif mudah, karena proses penentuan support vector dapat dirumuskan dalam QP problem. Dengan demikian jika kita memiliki library untuk menyelesaikan QP problem, dengan sendirinya SVM dapat diimplementasikan dengan mudah. Selain itu dapat diselesaikan dengan metode sekuensial sebagaimana penjelasan sebelumnya Kekurangan

Dari banyaknya kelebihan diatas SVM juga mempunyai banyak kekurangan antara lain

1. Sulit dipakai dalam problem berskala besar. Skala besar dalam hal ini dimaksudkan dengan jumlah sample yang diolah.

2. SVM secara teoritik dikembangkan untuk problem klasifikasi dengan dua class. Dewasa ini SVM telah dimodifikasi agar dapat menyelesaikan masalah dengan class lebih dari dua, antara lain strategi One versus rest dan strategi Tree Structure.

* + 1. **Naïve Bayes**

Naïve Bayes Classifier merupakan sebuah metoda klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metoda probabilitas dan statistic yang dikemukakan oleh ilmuan inggris Thomas Bayes, yaitu memprediksi peluang masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dari Naïve Bayes classifier ini adalah asumsi yang sangat kuat (naïf) akan indepensi dari masing-masing kondisi/kejadian.

Menurut Olson Delen (2008) menjelaskan Naïve Bayes untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vector informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlihat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari “master” table keputusan.

Naïve Bayes Classifier bekerja sangat baik dibandingkan dengan model classifier lainnya. Hal ini dibuktikan oleh Xhemali, Hinde Stone dalam jurnalnya “Naïve Bayes vs Decision Trees vs Neural Networks in the Classification of Training Web Pages” yang mengatakan bahwa “Naïve Bayes Classifier memiliki tingkat akurasi yang lebih baik disbanding model classifier lainnya”.

Keuntungan penggunakan metode Naïve Bayes Classifier adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variable independen, maka hanya varians dari suatu varibel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarian.

* + 1. **Logistic Regression**

Regresi logistik adalah sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear atau yang biasa disebut dengan istilah *Ordinary Least Squares (OLS) regression*. Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel terikat yang berskala dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah.

**Analisis regresi logistik** digunakan untuk menjelaskan hubungan antara variabel respon yang berupa data dikotomik/biner dengan variabel bebas yang berupa data berskala interval dan atau kategorik (Hosmer dan Lemeshow, 1989). Variabel yang **dikotomik/biner** adalah variabel yang hanya mempunyai dua kategori saja, yaitu kategori yang menyatakan kejadian sukses (Y=1) dan kategori yang menyatakan kejadian gagal (Y=0). pada model model linear umum komponen acak tidak harus mengikuti sebaran normal, tapi harus masuk dalam**sebaran keluarga eksponensial**. **Sebaran bernoulli** termasuk dalam salah satu dari sebaran keluarga eksponensial. Variabel respon Y ini, diasumsikan mengikuti distribusi Bernoulli.

Metode untuk mengestimasi parameter-parameter yang tidak diketahui dalam model regresi logistik ada 3 yaitu:

1. **Metode kemungkinan maksimum** (*Maximum Likelihood Method*).

2. **Metode kuadrat terkecil tertimbang noniterasi** (*Noniterative Weight Least Square Method*).

3. **Analisis fungsi diskriminan** (*Discriminant Fuction Analysis*).

Pada dasarnya metode maksimum Likelihood merupakan metode kuadrat terkecil tertimbang dengan beberapa proses iterasi, sedangkan metode noniterative weight least square method hanya menggunakan satu kali iterasi. kedua metode ini *asymptoticaly equivalent*, artinya jika ukuran sampel besar keduanya akan menghasilkan estimator yang identik. Penggunaan fungsi diskriminan mensyaratkan variabel penjelas yang kuantitatif berdistribusi normal. Oleh karena itu, penduga dari fungsi diskriminan akan*over estimate* bila variabel penjelas tidak berdistribusi normal.

Dari Ketiga metode di atas, metode yang banyak digunakan adalah metode **maksimum likelihood**dengan alasan lebih praktis (Nachrowi dan Usman, 2002). Metode maksimum likelihoood ini menduga parameter dengan nilai yang memaksimumkan fungsi likelihood (likelihood function).

* + 1. **MyPersonality**

MyPersonality merupakan project yang diciptakan oleh Michal Kosinki (Kosinki, 2015). Aplikasi ini digunakan oleh pengguna Facebook untuk mengisi kuesioner-kuesioner tentang diri mereka, kemudian memberikan hasil dari kuesioner tersebut yang berupa sifat mereka berdasarkan model kepribadian BIG 5 Personality Traits.

User-user yang telah menggunakan program ini kemudian di data profil nya dan dikumpulkan menjadi dataset yang telah digunakan oleh banyak peneliti di dunia.

* + 1. **Phyton**

Phyton dikembangkan oleh Guido van Rossum pada tahun 1990 di CWI, Amsterdam sebagai kelanjutan dari bahasa pemograman ABC.

Saat ini pengembangan Python terus dilakukan oleh sekumpulan pemrogram yang dikoordinir Guido dan Python Software Foundation. Python Software Foundation adalah sebuah organisasi non-profit yang dibentuk sebagai pemegang hak cipta intelektual Python sejak versi 2.1 dan dengan demikian mencegah Python dimiliki oleh perusahaan komersial. Saat ini distribusi Python sudah mencapai versi 2.6.1 dan versi 3.6.

Nama Python dipilih oleh Guido sebagai nama bahasa ciptaannya karena kecintaan guido pada acara televisi Monty Python s Flying Circus. Oleh karena itu seringkali ungkapan-ungkapan khas dari acara tersebut seringkali muncul dalam korespondensi antar pengguna Python.

Sisi utama yang membedakan Python dengan bahasa lain adalah dalam hal aturan penulisan kode program. Bagi para programmer di luar python siap-siap dibingungkan dengan aturan indentasi, tipe data, tuple, dan dictionary. Python memiliki kelebihan tersendiri dibandingkan dengan bahasa lain terutama dalam hal penanganan modul, ini yang membuat beberapa programmer menyukai python. Selain itu python merupakan salah satu produk yang opensource, free, dan multiplatform.

Beberapa fitur yang dimiliki Python adalah:

* memiliki kepustakaan yang luas; dalam distribusi Python telah disediakan modul-modul siap pakai untuk berbagai keperluan.
* memiliki tata bahasa yang jernih dan mudah dipelajari.
* memiliki aturan layout kode sumber yang memudahkan pengecekan, pembacaan kembali dan penulisan ulang kode sumber.
* berorientasi obyek.
* memiliki sistem pengelolaan memori otomatis (garbage collection, seperti java)
* modular, mudah dikembangkan dengan menciptakan modul-modul baru; modul-modul tersebut
* dapat dibangun dengan bahasa Python maupun C/C++.
* memiliki fasilitas pengumpulan sampah otomatis, seperti halnya pada bahasa pemrograman Java,
* python memiliki fasilitas pengaturan penggunaan ingatan komputer sehingga para pemrogram tidak perlu melakukan pengaturan ingatan komputer secara langsung.
  + 1. **Eclipse IDE**

Menurut Vogel dan Arthorne (2015), Eclipse dikenal sebagai Integrated Development Environment (IDE) untuk Java. Eclipse pun menguasai pasar global sebesar 65% sebagai IDE untuk Java programming.

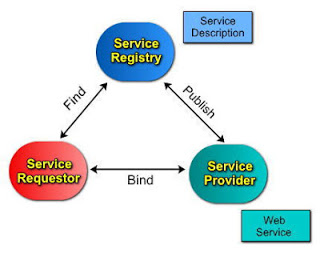
IDE Eclipse dapat diintegrasikan dengan komponen software tambahan, yang biasa disebut dengan plug-in dan beberapa plug-in dapat dikelompokkan menjadi features. Beberapa proyek dan perusahaan telah meningkatkan pemakaian IDE Eclipse untuk digunakan sebagai dasar untuk membuat sebuah aplikasi. Aplikasi ini dikenal sebagai Eclipse Rich Client Platform (Eclipse RCP). IDE Eclipse yang digunakan pada penelitian ini ialah Eclipse Neon (versi 4.6).

* + 1. **Web Service**

Web service adalah aplikasi sekumpulan data (database), perangkat lunak (software) atau bagian dari perangkat lunak yang dapat diakses secara remote oleh berbagai piranti dengan sebuah perantara tertentu. Secara umum,web service dapat diidentifikasikan dengan menggunakan URL seperti hanya web pada umumnya. Namun yang membedakan web service dengan web pada umumnya adalah interaksi yang diberikan oleh web service. Berbeda dengan URL web pada umumnya, URL web service hanya menggandung kumpulan informasi, perintah, konfigurasi atau sintaks yang berguna membangun sebuah fungsi-fungsi tertentu dari aplikasi.

Web service dapat diartikan juga sebuah metode pertukaran data, tanpa memperhatikan dimana sebuahdatabase ditanamkan, dibuat dalam bahasa apa sebuah aplikasi yang mengkonsumsi data, dan di platform apa sebuah data itu dikonsumsi. Web service mampu menunjang interoperabilitas. Sehingga web service mampu menjadi sebuah jembatan penghubung antara berbagai sistem yang ada.

Menurut W3C Web services Architecture Working Group pengertian Web service adalah sebuah sistem softwareyang di desain untuk mendukung interoperabilitas interaksi mesin ke mesin melalui sebuah jaringan. Interfaceweb service dideskripsikan dengan menggunakan format yang mampu diproses oleh mesin (khususnya WSDL). Sistem lain yang akan berinteraksi dengan web service hanya memerlukan SOAP, yang biasanya disampaikan dengan HTTP dan XML sehingga mempunyai korelasi dengan standar Web (Web Services Architecture Working Group, 2004).



Gambar 2.1 Bentuk Arsitektur Web Service

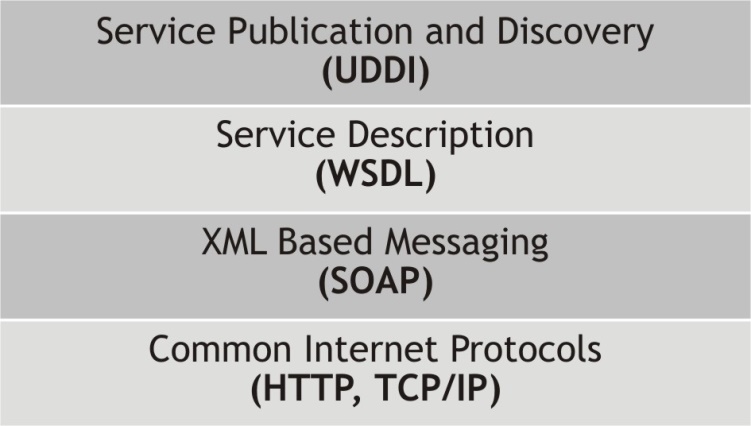
(Sumber : <http://www.ibm.com/developerworks/library/ws-wsilover/WebServicesArchitecture.jpg>)

Bagian-bagian dari entitas arsitektur web service adalah sebagai berikut:

* Service Provider: Berfungsi untuk menyediakan layanan/service dan mengolah sebuah registry agar layanan-layanan tersebut dapat tersedia.
* Service Registry: Berfungsi sebagai lokasi central yang mendeskripsikan semua layanan/service yang telah di-register.
* Service Requestor: Peminta layanan yang mencari dan menemukan layanan yang dibutuhkan serta menggunakan layanan tersebut.

Web pada umumnya digunakan untuk melakukan respon dan request yang dilakukan antara client dan server. Sebagai contoh, seorang pengguna layanan web tertentu mengetikan alamat url web untuk membentuk sebuahrequest. Request akan sampai pada server, diolah dan kemudian disajikan dalam bentuk sebuah respon. Dengan singkat kata terjadilah hubungan client-server secara sederhana.

Sedangkan pada web service hubungan antara client dan server tidak terjadi secara langsung. Hubungan antaraclient dan server dijembatani oleh file web service dalam format tertentu. Sehingga akses terhadap databaseakan ditanggani tidak secara langsung oleh server, melainkan melalui perantara yang disebut sebagai web service. Peran dari web service ini akan mempermudah distribusi sekaligus integrasi database yang tersebar di beberapa server sekaligus.



Gambar 2.2 Layer dari komponen-komponen Web Service

(Sumber : http://files.ekowins.webnode.com/200000001-856f5864af/komponen.jpg)

Komponen-komponen Web Service:

1. Layer 1: Protokol internet standar seperti HTTP dan TCP/IP.
2. Layer 2: Simple Object Access Protocol (SOAP), adalah sebuah XML-based mark-up language untuk pergantian pesan diantara aplikasi-aplikasi. SOAP berguna seperti sebuah amplop yang digunakan untuk pertukaran data object didalam network. SOAP mendefinisikan empat aspek didalam komunikasi: Message envelope, Encoding, RPC call convention, dan bagaimana menyatukan sebuah message didalam protokol transport.
3. Layer 3: Web Service Definition Language (WSDL), adalah sebuah XML-based language untuk mendeskripsikan XML. Ia menyediakan service yang mendeskripsikan service request dengan menggunakan protokol-protokol yang berbeda dan juga encoding. Ia akan memfasilitasi komunikasi antar aplikasi. WSDL akan mendeskripsikan apa yang akan dilakukan oleh web service, bagaimana menemukannya dan bagaimana untuk mengoperasikannya.
4. Layer 4: Universal Description Discovery and Integration (UDDI), adalah sebuah service registry bagi pengalokasian web service. UDDI mengkombinasikan SOAP dan WSDL untuk pembentukan sebuah registry API bagi pendaftaran dan pengenalan service. Ia menyediakan sebuah area umum dimana sebuah organisasi dapat mengiklankan keberadaan mereka dan service yang mereka berikan (web service). UDDI adalah sebuah framework yang mendefinisikan sebuah XML-based registry dimana sebuah organisasi dapat meng-upload informasi mengenai service yang mereka berikan. XML-based registry berisi nama-nama dari organisasi tsb, beserta service dan deskripsi dari service yang mereka berikan.
   * 1. **The Big Five Traits**

Kepribadian adalah pemikiran, emosi, dan perilaku tertentu yang menjadi ciri dari seseorang dalam menghadapi dunianya. Menurut Allport (2006) kepribadian merupakan sebuah organisasi dinamis di dalam psikis dan fisik individu yang menentukan karakteristik perilaku dan pikirannya. Menurut pandangan ini, setiap orang memiliki kualitas dasar yang unik.

Big Five merupakan pendekatan yang digunakan untuk melihat kepribadian manusia melalui trait yang tersusun dalam lima buah domain kepribadian yang telah dibentuk dengan menggunakan analisis faktor.

Dimensi Big Five kebanyakan berasal dari pendekatan leksikal (bahasa) terhadap trait. Dengan kata lain, orang mendeskripsikan, menguji, dan mengkategorisasikan orang lain, dan peringkat yang dihasilkan disederhanakan ke dalam lima dimensi. Trait-trait dari Big Five adalah sebagai berikut :

1. Extraversion (E)

Orang yang tinggi pada dimensi ini cenderung penuh semangat, antusias, dominan, ramah dan komunikatif. Orang yang sebaliknya akan cenderung pemalu, tidak percaya diri, submisif, dan pendiam.

1. Agreeableness (A)

Orang yang tinggi pada dimensi agreeableness cenderung ramah, kooperatif, mudah percaya dan hangat. Orang yang rendah dalam dimensi ini cenderung dingin, konfrontatif dan kejam.

1. Neuroticism (N)

Orang yang tinggi dalam dimensi Neuroticism biasanya gugup, sensitif, tegang dan mudah cemas. Orang yang rendah dalam dimensi ini cenderung tenang dan santai.

1. Conscientiousness

Orang yang tinggi dalam dimensi Conscientiousness umumnya berhati-hati, dapat diandalkan, teratur, dan bertanggung jawab. Orang yang rendah dalam dimensi ini cenderung ceroboh, berantakan dan biasanya tidak dapat diandalkan.

1. Openness (O)

Orang yang tinggi dalam dimensi Openness umumnya terlihat imajinatif, ceria, menyenangkan, kreatif dan artistic. Sebaliknya, orang yang rendah dalam dimensi ini umumnya dangkal, membosankan dan sederhana.

* + 1. **LIWC**

Linguistic Inquiry and Word Count (LIWC) adalah program penghitung jumlah kata yang merujuk kepada kamus gramatikal, psikologikal dan kategori content kata (Pennebaker et al, 2007). LIWC telah digunakan secara efisien untuk mengklasifikasikan kata beserta dimensi psikologikalnya untuk memprediksi hasil kepribadian dari seseorang. Membuat LIWC ini menjadi alat analisis teks yang digunakan secara luas dalam ilmu social sciences.

LIWC memiliki kamusnya tersendiri sebagai inti dari program ini. Versi saat ini memiliki 80 kategori dengan kurang lebih 4000 kata di dalamnya.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Category** | **Abbrev** | **Examples** | **Words in category** | **Internal Consistency (Uncorrected α)** | **Internal Consistency (Corrected α)** |
| Word count | WC | ­ | ­ | ­ | ­ |
| **Summary Language Variables** |  |  |  |  |  |
| Analytical thinking | Analytic | ­ | ­ | ­ | ­ |
| Clout | Clout | ­ | ­ | ­ | ­ |
| Authentic | Authentic | ­ | ­ | ­ | ­ |
| Emotional tone | Tone | ­ | ­ | ­ | ­ |
| Words/sentence | WPS | ­ | ­ | ­ | ­ |
| Words > 6 letters | Sixltr | ­ | ­ | ­ | ­ |
| Dictionary words | Dic | ­ | ­ | ­ | ­ |
| **Linguistic Dimensions** |  |  |  |  |  |
| Total function words | funct | it, to, no, very | 491 | .05 | .24 |
| Total pronouns | pronoun | I, them, itself | 153 | .25 | .67 |
| Personal pronouns | ppron | I, them, her | 93 | .20 | .61 |
| 1st pers singular | i | I, me, mine | 24 | .41 | .81 |
| 1st pers plural | we | we, us, our | 12 | .43 | .82 |
| 2nd person | you | you, your, thou | 30 | .28 | .70 |
| 3rd pers singular | shehe | she, her, him | 17 | .49 | .85 |
| 3rd pers plural | they | they, their, they’d | 11 | .37 | .78 |
| Impersonal pronouns | ipron | it, it’s, those | 59 | .28 | .71 |
| Articles | article | a, an, the | 3 | .05 | .23 |
| Prepositions | prep | to, with, above | 74 | .04 | .18 |
| Auxiliary verbs | auxverb | am, will, have | 141 | .16 | .54 |
| Common Adverbs | adverb | very, really | 140 | .43 | .82 |
| Conjunctions | conj | and, but, whereas | 43 | .14 | .50 |
| Negations | negate | no, not, never | 62 | .29 | .71 |
| **Other Grammar** |  |  |  |  |  |
| Common verbs | verb | eat, come, carry | 1000 | .05 | .23 |
| Common adjectives | adj | free, happy, long | 764 | .04 | .19 |
| Comparisons | compare | greater, best, after | 317 | .08 | .35 |
| Interrogatives | interrog | how, when, what | 48 | .18 | .57 |
| Numbers | number | second, thousand | 36 | .45 | .83 |
| Quantifiers | quant | few, many, much | 77 | .23 | .64 |
| **Psychological Processes** |  |  |  |  |  |
| Affective processes | affect | happy, cried | 1393 | .18 | .57 |
| Positive emotion | posemo | love, nice, sweet | 620 | .23 | .64 |
| Negative emotion | negemo | hurt, ugly, nasty | 744 | .17 | .55 |
| Anxiety | anx | worried, fearful | 116 | .31 | .73 |
| Anger | anger | hate, kill, annoyed | 230 | .16 | .53 |
| Sadness | sad | crying, grief, sad | 136 | .28 | .70 |
| Social processes | social | mate, talk, they | 756 | .51 | .86 |
| Family | family | daughter, dad, aunt | 118 | .55 | .88 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Friends | friend | buddy, neighbor | 95 | .20 | .60 |
| Female references | female | girl, her, mom | 124 | .53 | .87 |
| Male references | male | boy, his, dad | 116 | .52 | .87 |
| Cognitive processes | cogproc | cause, know, ought | 797 | .65 | .92 |
| Insight | insight | think, know | 259 | .47 | .84 |
| Causation | cause | because, effect | 135 | .26 | .67 |
| Discrepancy | discrep | should, would | 83 | .34 | .76 |
| Tentative | tentat | maybe, perhaps | 178 | .44 | .83 |
| Certainty | certain | always, never | 113 | .31 | .73 |
| Differentiation | differ | hasn’t, but, else | 81 | .38 | .78 |
| Perceptual processes | percept | look, heard, feeling | 436 | .17 | .55 |
| See | see | view, saw, seen | 126 | .46 | .84 |
| Hear | hear | listen, hearing | 93 | .27 | .69 |
| Feel | feel | feels, touch | 128 | .24 | .65 |
| Biological processes | bio | eat, blood, pain | 748 | .29 | .71 |
| Body | body | cheek, hands, spit | 215 | .52 | .87 |
| Health | health | clinic, flu, pill | 294 | .09 | .37 |
| Sexual | sexual | horny, love, incest | 131 | .37 | .78 |
| Ingestion | ingest | dish, eat, pizza | 184 | .67 | .92 |
| Drives | drives |  | 1103 | .39 | .80 |
| Affiliation | affiliation | ally, friend, social | 248 | .40 | .80 |
| Achievement | achieve | win, success, better | 213 | .41 | .81 |
| Power | power | superior, bully | 518 | .35 | .76 |
| Reward | reward | take, prize, benefit | 120 | .27 | .69 |
| Risk | risk | danger, doubt | 103 | .26 | .68 |
| Time orientations | TimeOrient |  |  |  |  |
| Past focus | focuspast | ago, did, talked | 341 | .23 | .64 |
| Present focus | focuspresent | today, is, now | 424 | .24 | .66 |
| Future focus | focusfuture | may, will, soon | 97 | .26 | .68 |
| Relativity | relativ | area, bend, exit | 974 | .50 | .86 |
| Motion | motion | arrive, car, go | 325 | .36 | .77 |
| Space | space | down, in, thin | 360 | .45 | .83 |
| Time | time | end, until, season | 310 | .39 | .79 |
| Personal concerns |  |  |  |  |  |
| Work | work | job, majors, xerox | 444 | .69 | .93 |
| Leisure | leisure | cook, chat, movie | 296 | .50 | .86 |
| Home | home | kitchen, landlord | 100 | .46 | .83 |
| Money | money | audit, cash, owe | 226 | .60 | .90 |
| Religion | relig | altar, church | 174 | .64 | .91 |
| Death | death | bury, coffin, kill | 74 | .39 | .79 |
| Informal language | informal |  | 380 | .46 | .84 |
| Swear words | swear | fuck, damn, shit | 131 | .45 | .83 |
| Netspeak | netspeak | btw, lol, thx | 209 | .42 | .82 |
| Assent | assent | agree, OK, yes | 36 | .10 | .39 |
| Nonfluencies | nonflu | er, hm, umm | 19 | .27 | .69 |
| Fillers | filler | Imean, youknow | 14 | .06 | .27 |

Tabel 2.1 Table LIWC2015 Output Variable Information from LIWC2015 Language Manual (Sumber : Pennebaker et al., 2015)

* + 1. **Splice**

Structured Programming for Linguistic Cue Extraction (SPLICE) adalah alat analisis linguistic yang dibangun oleh Kevin Moffit dan Justin Scott Giboney dari Universitas Arizona, Amerika Serikat (Moffit & Giboney, 2012). SPLICE merupakan tool yang cukup baru dan telah digunakan dalam berbagai penelitian yang berhubungan dengan NLP atau linguistic. SPLICE juga dapat diakses melalui web to membantu pengembangan aplikasi yang dapat terbantu dengan kekuatan teknik pemrosesan bahasa natural yang mumpuni. Berikut adalah fitur-fitur yang menjadi kategori kamus SPLICE:

1. Quantity ( 5 features )

Cues based on frequencies

1. Part of Speech ( 8 features )

All linguistic cues in the Parts of Speech category are calculated using a part of speech tagger based on the Brown corpus.

1. Immediacy ( 2 features )

Cues that indicate immediacy

1. Pronouns ( 10 features )

Cues that count pronoun frequencies

1. Positive Self Evaluation ( 3 features )

Cues that relate to the positive self evaluation of a speaker.

1. Negative Self Evaluation ( 3 features )

Cues that relate to the negative self evaluation of a speaker.

1. Influence ( 8 features )

Cues that indicate the speaker is attempting to influence someone.

1. Deference ( 4 features )

Cues that indicate the speaker is accommodating outside influence.

1. Whissel ( 3 features )

Cues that count returns scores based on the Whissel Dictionary of Affect in Language

1. Complexity ( 12 features )

Cues that relate to text complexity

1. Spoken Word ( 18 features )

Cues that relate to a speaking or writing style

1. Tense ( 4 features )

Counts of verbs of a specific tense

1. Sentiwordnet ( 3 features )

Calculates positivity, negativity, and objectivity based on Sentiwordnet

1. Readability ( 11 features )

Cues that return readability scores; readability scores

* 1. **Penelitian Terkait**

Berbagai penelitian mengenai prediksi atau hubungan antara kepribadian berbasis model “Big 5 Traits” telah banyak dilakukan sebelumnya. Berbagai penelitian tersebut juga menggunakan data dari berbagai sosial media, salah satunya Facebook yang merupakan target dari penelitian kita. Dari penelitian-penelitian tersebut, terdapat perkembangan dan kemajuan dalam akurasi menentukan sifat dari pengguna berdasarkan berbagai faktor dan fitur. Tidak hanya dari segi linguistik, beberapa peneliti juga mulai menggunakan fitur lainnya seperti umur, gender, jumlah foto, jumlah likes dan sebagainya. Berikut adalah beberapa penelitian prediksi kepribadia yang telah dilakukan para peneliti sebelumnya:

* + 1. **Personality and Patterns of Facebook Usage**

Kesuksesan seorang individu bergantung terhadap apa yang ditampilkan oleh dirinya kepada orang lain. Kesuksesan di lingkungan kerja, percintaan dan mendapat dukungan positif dari seseorang sangatlah bergantung terhadap apa yang orang lain pikirkan tentang diri kita. Interaksi manusia, sosialisasi dan komunikasi di dalam dunia online yang semakin berkembang berarti menunjukkan pandangan diri kita terhadap orang lain juga semakin penting.

Salah satu sosial media yang paling berkembang, Facebook, telah mempengaruhi sebagian besar populasi manusia di dunia. Sekitar 800 juta pengguna menggunakan Facebook 40 menit setiap hari. Profil Facebook menjadi sumber informasi penting bagi kita untuk menciptakan pandangan kita terhadap orang lain. Dari beberapa penelitian sebelumnya, menunjukkan bahwa kepribadian seseorang dapat diprediksi oleh orang lain berdasarkan profil dari Facebook mereka.

Penelitian ini berfokus pada bagaimana prediksi kepribadian dapat diambil melalui fitur-fitur berbeda dari profil Facebook. Peneliti mengekstraksi high-level features dari profil Facebook dan menunjukkan korelasinya dengan kepribadian pengguna, sesuai dengan kuesioner model kepribadian Five Factor Model. Profil fitur yang digunakan dalam penelitian ini dibagi ke dalam dua aspek yaitu,

1. Aspek profil yang bergantung sepenuhnya terhadap kegiatan dari pengguna itu sendiri yang meliputi: Jumlah foto yang di upload, jumlah event yang diikuti dan jumlah grup yang dibentuk, serta jumlah objek atau foto yang di like oleh user

2. Aspek profile yang bergantung kepada kegiatan pengguna dan teman dari si pengguna yang meliputi: Jumlah berapa kali pengguna di tag dalam foto dan ukuran pertemanan si pengguna (number of friends or network size).

|  |  |
| --- | --- |
| Feature | Details |
| Friends | number of Facebook friends |
| Groups | number of associations with groups |
| Likes | number of Facebook “likes” |
| Photos | number of photos uploaded by user |
| Statuses | number of status updates by user |
| Tags | number of times others “tagged” user in photos |

Table 1. Fitur Profil Facebook yang digunakan dalam penelitian ini

Penelitian ini menggunakan Five Factor Model yang saat ini paling banyak digunakan dan diterima, dimana model untuk menentukan kepribadian manusia ini telah sangat baik dipelajari. Beberapa penelitian sebelumnya menunjukkan bahwa kepribadian berkorelasi dengan banyak aspek dalam kehidupan, termasuk kesuksesan dalam dunia kerja, daya tarik, keharmonisan pernikahan dan kebahagiaan. Penelitian sebelumnya telah menunjukkan kepribadian berkorelasi dengan total penggunaan internet, sosial media dan situs jaringan sosial lainnya. Tetapi, penelitian ini lebih berfokus pada jumlah berapa kali seorang pengguna menggunakan alat-alat atau fitur tersebut daripada bagaimana pengguna menggunakan alat atau fitur tersebut.

Beberapa hipotesis telah didapatkan dari hubungan antara kepribadian dan fitur profil Facebook. Ross et al. (2009), mempelajari mengenai hubungan antara kepribadian dan penggunaan jaringan sosial. Mereka menghasilkan beberapa hipotesis antara lain:

1. Relasi positif antara Extraversion dengan penggunaan Facebook, jumlah teman di Facebook dan jumlah grup si pengguna.
2. Relasi Positif antara Neuroticism dengan informasi privasi yang dimunculkan di Facebook.
3. Relasi Positif antara Agreeableness dengan jumlah teman di Facebook.
4. Relasi positif antara Openness dan banyaknya jumlah fitur Facebook yang dipakai.
5. Relasi negative antara Conscientiousness and penggunaan Facebook secara keseluruhan.

Sayangnya penelitian di atas menggunakan dataset yang relative cukup kecil (n = 97) dan sampel homogen (sebagian besar murid perempuan yang mengambil mata kuliah sama di universitas yang sama) yang membuat kurangnya kekuatan analisis dan sulitnya diterapkan kepada populasi pada umumnya.

Beberapa penelitian sebelumnya menggunakan dataset atau sampel yang terbatas dan homogen sehingga menghasilkan sesuatu yang kontradiksi. Tujuan utama penelitian ini adalah menggunakan sampel pengguna yang besar dan representatif untuk menjawab pertanyaan bagaimana profil Facebook dapat menentukan kepribadian. Peneliti mengemukakan hipotesis sebagai berikut:

1. Openness dan Neuroticism memiliki korelasi positif dengan jumlah update status, foto, grup dan jumlah likes pengguna.
2. Conscientiousness memiliki korelasi negatif dengan semua aspek dari penggunaan Facebook.
3. Extraversion memiliki korelasi positif dengan semua aspek dari penggunaan Facebook.
4. Agreeableness memiliki korelasi positif dengan jumlah teman, grup dan jumlah likes pengguna.