ANALISIS SENTIMEN MENGGUNAKAN *TEXT MINING* DENGAN METODE *NAÏVE BAYES* DAN REGRESI LOGISTIK

ABDUL GAFUR DARUSSALAM

H 121 15 002

Pembimbing Utama	: Andi Kresna Jaya S.Si., M.Si
Pembimbing Pendamping	: Sri Astuti Thamrin S.Si., M.Stat., Ph.D
Penguji	: Dr. Dr. Georgina Maria Tinungki, M.Si
	Dr. Anna Islamiyati, S.Si., M.Si

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2021

ANALISIS SENTIMEN MENGGUNAKAN TEXT MINING DENGAN METODE NAÏVE BAYES DAN REGRESI LOGISTIK

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin Makassar

ABDUL GAFUR DARUSSALAM

H 121 15 002

PROGRAM STUDI STATISTIKA DEPARTEMEN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2021

LEMBAR PERNYATAAN KEOTENTIKAN

Saya yang bertanda tangan di bawah ini menyatakan dengan sungguh-sungguh bahwa skripsi yang saya buat dengan judul:

Analisis Sentimen Menggunakan Text Mining Dengan Metode Naïve Bayes Dan Regresi Logistik

adalah benar hasil karya saya sendiri, bukan hasil plagiat dan belum pernah dipublikasikan dalam bentuk apapun.

Makassar, 10 Mei 2021

ABDUL GAFUR DARUSSALAM

NIM. H 121 15 002

ANALISIS SENTIMEN MENGGUNAKAN TEXT MINING DENGAN METODE NAÏVE BAYES DAN REGRESI LOGISTIK

Disetujui Oleh:

Pembimbing Utama	Pembimbing Pertama
Andi Kresna Jaya, S.Si., M.Si	Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.
NIP. 19731228 200003 1 001	NIP. 19740713 199903 2 001

Ketua Departemen Statistika

Dr. Nurtiti Sunusi, S.Si., M.Si.
NIP. 19720117 199703 2 002

Pada Tanggal: 10 Mei 2021

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama : Abdul Gafur Darussalam

NIM : H 121 15 002

Program Studi : Statistika

Judul Skripsi : Analisis Sentimen Menggunakan Text Mining Dengan Metode Naïve Bayes Dan Regresi Logistik

Telah berhasil dipertahankan di hadapan dewan penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Sains pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

DEWAN PENGUJI

		Tanda Tangan
1. Ketua	: Andi Kresna Jaya, S.Si., M.Si	(.....)
2. Sekretaris	: Sri Astuti Thamrin, S.Si., M.Stat., Ph.D.	(.....)
3. Anggota	: Dr. Dr. Georgina Maria Tinungki, M.Si.	(.....)
4. Anggota	: Dr. Anna Islamiyati, S.Si., M.Si.	(.....)

Ditetapkan di : Makassar

Tanggal : 10 Mei 2021

KATA PENGANTAR

Segala puji bagi Allah *Subhanahu Wa ta'ala*, Tuhan atas langit dan bumi beserta segala isinya. Karena, berkat nikmat dan karuniaNYA sehingga penulisan skripsi ini dapat terselesaikan. Shalawat serta salam semoga senantiasa tercurahakan kepada Baginda Rasulullah Muhammad *Shallallahu Alaihi Wasallam* dan kepada para keluarga serta sahabat beliau, yang senantiasa menjadi teladan yang baik.

Alhamdulillah, skripsi dengan judul "Analisis Sentimen Menggunakan Text Mining Dengan Metode Naïve Bayes Dan Regresi Logistik" yang disusun sebagai salah satu syarat akademik untuk meraih gelar Sarjana Sains pada Program Studi Statistika Fakutas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin ini dapat dirampungkan. Tentunya, dalam penulisan skripsi ini, penulis mampu melewati berbagai hambatan dan masalah berkat bantuan moril dan materiil, serta dorongan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang tak terhingga kepada orang tua penulis, Ibunda **Indo Upe** dan Almarhum Ayahanda **Baharuddin** , sebagai tempat kembali setelah pergi, dan tempat terlelap dikala lelah, terima kasih atas kasih sayang, doa, dan nasihat yang tulus sebagai bekal kehidupan serta motivasi dan bimbingan yang besar juga untuk saudara-saudara saya **Misbahuddin, Sustiana dan Nur Aeni** .

Penghargaan dan ucapan terima kasih dengan penuh ketulusan juga penulis ucapkan kepada:

1. Ibu **Dr. Nurtiti, Sunusi, S.Si., M.Si.** selaku Ketua Departemen Statistika, **segenap dosen pengajar**, **staf jurusan Statistika**, dan **staf Fakultas MIPA**, yang telah membekali ilmu dan kemudahan-kemudahan kepada penulis dalam berbagai hal selama menjadi mahasiswa di jurusan Statistika.
2. Bapak **Andi Kresna Jaya, S.Si., M.Si.** sebagai dosen pembimbing utama sekaligus ketua tim penguji atas ilmu yang beliau berikan selama proses perkuliahan, dan kesediaan beliau dalam membimbing, serta memotivasi penulis baik diluar maupun dalam penyusunan skripsi ini.

3. Ibu **Sri Astuti Thamrin, M.Stat, Ph.D.** sebagai dosen pembimbing pertama sekaligus sekretaris tim penguji atas ilmu yang beliau berikan selama proses perkuliahan, dan bimbingan serta segala bentuk bantuan yang telah beliau berikan dalam penyusunan skripsi ini.
4. Ibu **Dr. Dr. Georgina Maria Tinungki, M.Si.** sebagai penasehat akademik dan anggota tim penguji atas segala ilmu dan bimbingan yang telah beliau berikan selama proses perkuliahan bentuk kritik dan masukan yang membangun selama proses penyusunan skripsi ini
5. Ibu **Dr. Anna Islamiyati, S.Si., M.Si.** selaku penguji yang selama seminar telah banyak memberikan kritik dan saran yang sangat berharga dalam perbaikan skripsi ke arah yang lebih baik.
6. Kanda **Siswanto, S.Si., M.Si.** atas bimbingan, motivasi dan bantuan yang diberikan kepada penulis selama proses perkuliahan.
7. **Yusnia Ahmad S.P** atas segala bentuk bantuan yang diberikan baik berupa moril dan materil selama proses perkuliahan dan dalam penyusunan skripsi, serta motivasi yang selalu diberikan kepada penulis.
8. **Masjidil Aqsha dan Waode Rahmalia Safitri.** atas segala bantuan yang diberikan, utamanya dalam proses penyusunan skripsi baik moril dan materil. Semoga dipermudah segala urusannya.
9. Sahabat sekaligus saudaraku **Muh. Anugrah Ariansyah, Andi Rahmat Kaswar, Wahyudi Bayu S.R, Yoris Rombe, Muhammad Aris, Nifal Gusti, Ricky Risman Ali, Nurul hidayanti, Tuty Awalia, Ainun Mutahharah dan Hesti Erdayanti** atas kebersamaan, kepedulian, suka-duka, canda tawa yang telah kita lewati selama ini. Semoga persahabatan kita yang telah terjalin tidak pernah usai.
10. Keluarga besar **Statistika 2015** atas segala bentuk dukungan dan bantuan selama proses perkuliahan. Terkhusus kepada teman seperjuangan dalam penyusunan skripsi,**Ihza Kurniawan, Muh. Fadil dan Ade Kurniawan** . Semoga kesuksesan selalu kita dapatkan dalam setiap langkah-langkah kita.
11. Keluarga besar **HIMASTAT Dan HIMATIKA** Terkhusus **Simetris 2015 Muh. Anwar Sadad dan Muh. Nur Khaliq** atas dedikasi serta semangatnya

dalam menjalankan roda-roda organisasi sekaligus proses akademik. Semoga ilmu yang didapatkan dapat diterapkan dalam kehidupan sehari-hari.

12. Keluarga Besar **KMF MIPA 2015** Terkhusus **Tampan Otodidak** atas jalinan saudara tak sedarah dalam menempuh masa-masa perkuliahan dan organisasi.
13. Seluruh pihak yang yang tidak dapat disebutkan satu per satu atas segala bentuk kontribusi, partisipasi, serta motivasi yang diberikan kepada penulis selama ini. Semoga apa yang kita berikan, dilipatgandakan oleh Tuhan Yang Maha Kaya.

Penulis menyadari bahwa masih banyak kekurangan dalam tugas akhir ini, untuk itu dengan segala kerendahan hati penulis memohon maaf. Akhir kata, semoga tulisan ini memberika manfaat untuk pembaca.

Makassar, 10 Mei 2021

Penulis

PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR
UNTUK KEPENTINGAN AKADEMIS

Sebagai civitas akademik Universitas Hasanuddin, saya yang bertanda tangan di bawah ini:

Nama	: Abdul Gafur Darussalam
NIM	: H 121 15 002
Program Studi	: Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam
Jenis Karya	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Hasanuddin **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas karya ilmiah saya yang berjudul:

"Analisis Sentimen Menggunakan Text Mining Dengan Metode Naïve Bayes Dan Regresi Logistik"

Beserta perangkat yang ada (jika diperlukan). Terkait dengan hal di atas, maka pihak universitas berhak menyimpan, mengalih-media/format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di Makassar, pada tanggal 10 Mei 2021

Yang menyatakan

(Abdul Gafur Darussalam)

ABSTRAK

Twitter menyediakan *Application Programming Interface Streaming* (APIs) untuk memfasilitasi data *crawling*. Akhir-akhir ini timbul keresahan masyarakat yang sempat menjadi trending topik di twitter mengenai berita Vaksin Covid-19. Sehingga perlu dilakukan penelitian untuk mengetahui tanggapan masyarakat berdasarkan sentimen di media sosial (twitter) terhadap Vaksin Covid-19. Teks mining merupakan proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah data tak terstruktur yang nantinya akan diperoleh pola-pola data, trend dan ekstraksi pengetahuan yang potensial dari data teks. Metode NBC telah banyak digunakan dalam penelitian mengenai text mining karena memiliki kelebihan yaitu algoritma sederhana tapi memiliki akurasi yang tinggi dan metode Regresi logistik memiliki interprestasi probabilistik yang bagus. Hasil ketepatan klasifikasi menggunakan Naïve Bayes Classifier pada data Vaksin Covid-19 diperoleh Akurasi. G-Mean dan AUC sebesar 63,75% ,61,21 dan 62,12% dan Regresi Logistik Biner pada data Vaksin Covid19 diperoleh Akurasi,G-Mean dan AUC sebesar 72,6% ,69,9 dan 70,8%.

Kata Kunci: Teks Mining, Vaksin Covid-19, Analsis Sentimen, Naïve-Bayes, Regresi Logistik Biner.

ABSTRACT

Twitter provides *Application Programming Interface Streaming* (APIs) to facilitate crawling data. Lately there has been public unrest which has become a trending topic on twitter regarding the news of the Covid-19 Vaccine. So it is necessary to do research to find out people's responses based on sentiment on social media (twitter) about the Covid-19 Vaccine. Text mining is the process of pattern extraction (useful information and knowledge) from a number of unstructured data which will later obtain data patterns, trends and potential knowledge extraction from text data. The NBC method has been widely used in research on text mining because it has the advantage of being a simple algorithm but has high accuracy and the logistic regression method has a good probabilistic interpretation. The results of classification accuracy using the NBC on Covid-19 Vaccine data obtained Accuracy, G-Mean and AUC of 63.75%, 61.21 and 62.12% and Binary Logistic Regression on Covid19 Vaccine data obtained Accuracy, G-Mean and AUC of 72.6%, 69.9 and 70.8%.

Keywords: Text Mining, Covid-19 Vaccine, Sentiment Analysis, NBC, Binary Logistic Regression.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Di era globalisasi, media social tidak bisa dipisahkan oleh kehidupan manusia. Sebut saja *twitter*, sebuah media baru berjenis *microblogging* yang bisa memberikan kita kemudahan untuk mendapatkan berita secara cepat dan singkat. Pengguna twitter di Indonesia menempati peringkat 5 terbesar di dunia dibawah USA, Brazil, Jepang, dan Inggris yaitu mencapai angka 19,5 juta pengguna twitter dari total 300 juta pengguna global (Kemenkominfo, 2016). Pengguna twitter dapat mengemukakan pendapatnya terhadap suatu produk atau mengomentari suatu masalah melalui tweet. Tweet pada setiap pengguna twitter dapat berpengaruh dalam pembentukan citra suatu produk atau program karena semakin banyak suatu topik tertentu diulas dalam tweet pengguna maka topik tersebut dapat menjadi trending topic di twitter (Kurniawan, 2017). Twitter telah menyediakan *Application Programming Interface* (API) yaitu sekumpulan fungsi atau protocol yang disediakan untuk pengguna dalam rangka mengembangkan sebuah aplikasi (Blanchette, 2008). Twitter API memungkinkan pengguna untuk mengakses dan mendapatkan informasi mengenai tweet, profil pengguna data follower dan lainnya. Hal tersebut menjadikan Twitter sebagai *microblog* yang banyak diminati perusahaan, organisasi, maupun individu dalam mendapatkan opini public mengenai suatu topik tertentu (Kurniawan, 2017).

Akhir-akhir ini timbul keresahan masyarakat yang sempat menjadi trending topik di twitter mengenai berita “Vaksin Covid19”. Salah satu upaya untuk menekan angka kasus Covid-19 yang kian meningkat adalah dengan penyediaan vaksin Covid-19 dari pemerintah. Meski dalam tahp uji klinis, keberadaan vaksin ini diharapkan dapat melindungi masyarakat Indonesia dari pendemi, meski demikian hingga saat ini, efekttvitas dan keamanan vaksin Covid-19 masih diteliti dalam tahap uji klinis oleh pemerintah dan berbagai lembaga terkait. Pro dan kontra

terkait vaksinisasi covid-19 ini masih terus tumbuh. Masih banyaknya masyarakat yang khawatir akan efektivitas dan efek samping dari vaksin tersebut lantaran ke simpang siuran pemberitaan yang terjadi di media social akan vaksin. Sehingga perlu dilakukan penelitian untuk mengetahui tanggapan masyarakat berdasarkan sentimen di media sosial (twitter) terhadap Vaksin Covid19.

Teks mining merupakan proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah data tak terstruktur yang nantinya akan diperoleh pola-pola data, trend dan ekstraksi pengetahuan yang potensial dari data teks (Turban, 2011). Salah satu tujuan penggunaan *teks mining* adalah analisis sentimen. Analisis sentimen atau disebut juga *opinion mining* merupakan proses memahami, mengestrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini terhadap sebuah masalah atau objek oleh seseorang apakah cenderung beropini negatif atau positif (Liu, 2012). Sebelum melakukan analisis sentimen, diperlukan praproses teks dengan metode *text mining* untuk mengolah data teks agar siap dianalisis (Kurniawan, 2017). Praproses teks tersebut meliputi *case folding*, tokenisasi, *stopword*, dan *steaming*.

Terdapat banyak metode klasifikasi dalam ilmu statistika yang digunakan untuk analisis sentimen seperti: Analisis Diskriminan, Regresi Logistik, *Naïve Bayes*, *Support vector Machine* (SVM) dan lain-lain, namun metode yang sering digunakan dalam klasifikasi teks adalah metode Regresi logistic dan *naive bayes classifier* (NBC). Metode Regresi Logistik Biner merupakan metode klasik yang digunakan untuk mengetahui pola hubungan antara variabel respon yang bersifat biner yakni terdiri dari 0 dan 1 dengan variabel prediktornya sedangkan Naive Bayes Classifier adalah suatu klasifikasi yang berdasarkan pada teorema Bayes yang bertujuan dalam menghitung peluang pada tiap kelas serta memiliki asumsi bahwa hubungan antar kelas adalah independen. Metode NBC telah banyak digunakan dalam penelitian mengenai *text mining* karena memiliki kelebihan yaitu algoritma sederhana tapi memiliki akurasi yang tinggi (Rish, 2006). Sedangkan metode Regresi logistik memiliki interprestasi *probabilistic* yang bagus (Pradesa,

2019).

Penelitian yang pernah dilakukan mengenai sentimen analisis dengan terlebih dahulu mengklasifikasian sentimen data awal secara manual adalah Analisis Sentimen Media Sosial (*Twitter*) Terhadap Layanan Provider Telkomsel Menggunakan Metode Multinomial Naïve Bayes. Penelitian tersebut mendapatkan hasil akurasi sebesar 81,25% (Yanti, 2018). Penelitian lain dilakukan Widhianingsih (2016) yang berjudul Aplikasi Text Mining untuk Automasi Klasifikasi Artikel dalam Majalah Online Wanita Menggunakan Naïve Bayes Classifier (NBC). Penelitian tersebut mengklasifikasi teks artikel majalah 4 wanita. Tingkat akurasi model NBC sebesar 80,71%.

Pada penelitian ini, struktur data yang digunakan terdiri dari variabel independen yaitu kata dasar tweet yang telah dilakukan praproses text dan variabel dependen yaitu klasifikasi sentimen tweet (Pro dan Kontra) secara manual terlebih dahulu. Lalu data yang diperoleh akan divisualisasikan secara geospasial untuk menunjukkan dari mana orang-orang paling banyak men-tweet mengenai Vaksin Covid-19 dan akan ditampilkan dalam warna yang berbeda sesuai dengan evaluasi sentimental yang kami peroleh. Misalnya, jika tweet bernilai negatif dalam evaluasi maka akan ditampilkan dalam penanda warna merah di peta dan hijau untuk positif. Dalam penelitian ini, peneliti juga ingin melakukan serta mendapatkan berapa hasil akurasi klasifikasi sentimen mengenai tanggapan masyarakat terhadap Vaksin Covi-19 menggunakan *Text Mining* data twitter dengan metode *Naïve Bayes Classifier* dan Regresi Logistik Biner serta ingin melihat asal daerah mana saja yang bersentimen Pro dan kontra mengenai omnibus law dengan visualisasi map.

Berdasarkan uraian diatas, maka penulis tertarik melakukan penelitian tugas akhir dengan judul **"Analisis Sentimen Menggunakan *Text Mining* Dengan Metode *Naïve Bayes* dan Regresi Logistik".**

1.2 Rumusan Masalah

Rumusan masalah yang dapat diambil dari Penelitian ini adalah :

1. Bagaimanakah persentase ketepatan klasifikasi sentimen yang didapatkan dengan menggunakan metode *Naïve Bayes Classifier* dan Regresi Logistik Biner pada pandangan masyarakat terhadap Vaksin Covid-19 ?

2. Bagaimana pemetaan secara spasial hasil tweet masyarakat mengenai Vaksin Covid-19?

1.3 Tujuan Penelitian

Dari rumusan masalah di atas, didapatkan tujuan sebagai berikut :

1. Mengetahui persentase ketepatan klasifikasi sentimen masyarakat dengan metode *Naïve Bayes* dan Regresi Logistik Biner pada kasus Vaksin Covid-19.
2. Mendapatkan visualisasi pemetaan pada masing-masing tweet sentiment.

1.4 Batasan Masalah

Batasan untuk penelitian ini adalah sebagai berikut:

1. Data yang diambil adalah tweet berupa kata "Vaksin Covid19" bersumber dari media sosial (*Twitter*) yang diunggah pada tanggal 28 Januari 2021 sampai 4 Februari 2021.
2. Data yang digunakan adalah data teks berbahasa Indonesia.

BAB 2

TINJAUAN PUSTAKA

2.1 *Text Mining*

Text mining adalah lintas disiplin ilmu yang mengacu pada pencarian informasi, *data mining*, *machine learning,* statistik, dan komputasi linguistik (Gusriani, 2016). *Text mining* juga dikenal sebagai data mining teks. *Text mining* dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam data mining (Nurhada, 2015).

Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan dalam *text mining* adalah sekumpulan teks. Pada dasarnya proses kerja dari *Text Mining* banyak mengadopsi dati penelitian data mining namun menjadi perbedaan adalah pola yang digunakan oleh *Text Mining* diambil dari sekumpulan Bahasa alami yabg tidak terstruktur sedangkan dalam *data mining* pola yang diambil dari database yang terstrukur. Tahap-tahap *Text Mining* secara umum adalah praproses teks dan *feature selection* (Kurniawan, 2017).

2.2 *Twitter Crawling*

Twitter menyediakan *Application Programming Interface Streaming* (APIs) untuk memfasilitasi data *crawling*. API memudahkan pengguna untuk mengambil data tweet secara real time. Tujuan awal dibentuknya *Twitter* API ini adalah untuk mengetahui relasi dan interaksi antara pengguna, namun sebaliknya *Twitter* API banyak digunakan untuk menggali informasi komunitas tertentu atas pandangannya terhadap topik yang sedang tren (Nguyen & Zheng, 2014).

Crawling adalah proses pengambilan sejumlah besar halaman web dengan cepat ke dalam suatu tempat penyimpanan local dan mengindeksnya berdasar sejumlah kata kunci (Liu, 2012). Mesin pencari web bekerja dengan cara

menyimpan informasi tentang banyak halaman web, yang diambil langsung dari situs dan untuk penelitian ini akan mengambil opini dari akun twitter tertentu.

2.3 Praproses Teks

Praproses teks merupakan tahapan awal dalam pengolahan teks yang digunakan untuk pengubahan bentuk dokumen menjadi data yang terstrukur sesuai kebutuhannya agar dapat diolah lebih lanjut dalam proses text mining. Tahapan praproses teks dalam klasifikasi bertujuan untuk meningkatkan akurasi klasifikasi data (Kurniawan, 2017). Berikut tahapan dalam praproses teks:

a) *Case folding* adalah proses penyamaan case dalam sebuah dokumen. Ini dilakukan untuk mempermudah pencarian. Tidak semua dokumen teks konsisten dalam penggunaan huruf capital. Oleh karena itu peran *case folding* dibutuhkan dalam menkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kecil).
b) Tokenisasi adalah proses untuk membagi teks yang berasal dari kalimat atau paragraph menjadi bagian-bagian tertentu (Manning dkk., 2009). Sebagai contoh, tokenisasi dari kalimat "Azizah senang sekali keliling Papua" menghasilkan lima token, yakni: "Azizah", "senang", "sekali", "keliling", "papua". Biasanya, yang menjadi acuan pemisah antar token adalah spasi dan tanda baca. Tokenisasi seringkali dipakai dalam ilmu linguistik dan hasil tokenisasi berguna untuk analisis teks lebih lanjut.
c) *Stopword* didefenisikan sebagai sebuah kata yang sangat sering muncul dalam suatu dokumen teks yang kurang memberikan arti pnting terhadap isi dokumen (Patel & Shah, 2013). Kata depan dan konjungsi merupakan kandidat besar dari daftar *stopword* yang harus dihilangkan. Untuk dokumen berbahasa Indonesia, contoh dari kata-kata penghubung adalah "yang", "di", "dan", "itu", "dengan". Langkah ini bermanfaat untuk mengurangi jumlah feature yang akan digunakan. Pada stopword ini juga peneliti mengabaikan emoji, sehingga peneliti befokus pada penelitian dengan fitur text karena emoji dapat menggangu proses analisis sentimen (Alita, 2018).

d) *Steamming* adalah Proses untuk mendapatkan kata dasar dengan cara menghilangkan awalan, akhiran, sisipan dan Confixes (kombinasi awalan dan akhiran (Kurniawan, 2017).
e) *Removal URL* yaitu Proses menghapus URL atau alamat website yang ada pada *tweet* (Kurniawan, 2017).

Penjelasan dari hasil simulasi Praproses teks pada data *tweet* sebagai berikut:

Contoh Tweet	@hendyaw Ayo ingat untuk bersyukur,vaksin siap cegah penularan Covid19
Menghapus Simbol RT, Username dan Url	Ayo ingat untuk bersyukur vaksin siap cegah penularan Covid19
Case Folding	ayo ingat untuk bersyukur vaksin siap cegah penularan covid19
Mengahpus Stopword	ayo ingat untuk syukur vaksin siap cegah nular covid19

Streamming dan Tokenisasi								
ayo	ingat	untuk	syukur	vaksin	siap	cegah	nular	covid19

Gambar 2.1 Simulasi Praproses Teks

Untuk *tweet* berikutnya "RT @taharuddin_id:Ayo Tolak Vaksin Covid19 ://t.co/WQWMhiINuO". Akan dilakukan praproses teks dengan langkah-langkah yang sama sehingga menghasilkan hasil praproses terakhir sebagai berikut:

ayo	tolak	Vaksin	Covid19

Gambar 2.2 Contoh hasil praproses teks

Dari kedua contoh hasil Praproses teks pada *tweet* diatas misalkan *tweet* pertama (Gambar 2.1) ialah sentimen positif dan *twee*t kedua (Gambar 2.2) ialah sentimen negatif, maka didapat struktur data setelah praproses teks sebagai berikut:

Tweet ke	Variabel Prediktor									
	ayo	ingat	untuk	syukur	vaksin	siap	cegah	nular	covid19	tolak
1	1	1,3	1,3	1,3	1	1	1	1	1	0
2	1	0	0	0	1	0	0	0	1	1,3

Gambar 2.3 Struktur data Hasil Praproses Teks

Pembentukan struktur data setelah dilakukan praproses teks seperti pada Gambar 2.3 yaitu menjadikan setiap kata menjadi variabel predictor dan meletakkannya pada satu baris. Jika terdapat tambahan kata (variabel predictor) dari *tweet* baru, maka kata tersebut diletakkan pada baris yang sama dan dikolom berikutnya. Namun jika terdapat kata yang sama atau kata yang telah ada pada struktur data, maka kata tersebut tidak dimasukkan lagi pada struktur data. Sehingga tidak terdapat kata atau variabel predictor yang sama dalam struktur data. Nilai dari setiap kata tersebut merupakan bobot yang diperoleh dari proses *TF-IDF* seperti yang terdapat pada Gambar 2.3 mengenai contoh struktur data setelah Praproses teks.

2.4 Sentimen Analisis

Opini menurut kamus besar Bahasa Indonesia sangat sederhana, yaitu pendapat, pikiran atau pendirian. Sedangkan sentimen adalah opini yang didalamnya mengandung perasaan atau emosi (Mulsy, 2015). Sentimen analisis atau biasa disebut *opinion mining* adalah bidang ilmu yang menganalisa pendapat, sentiment, evaluasi, penilaian, sikap dan emosi public terhadap entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa topik (Liu, 2012).

Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif atau negatif. Sentimen analisis juga dapat menyatakan perasaaan emosiaonal sedih, gembira atau marah. Kita juga dapat mencari pendapat tentang produk-produk, merek tau orang-orang dan menentukan apakah mereka dilihat positif atau negative di *web* (Kurniawan, 2017).

Ekspresi atau sentimen mengacu pada focus topik tertentu, pernyataan pada satu topik mungkin akan berbeda makna dengan pernyataan yang sama pada subjek yang berbeda. Oleh karena itu pada beberapa penelitian, pekerjaan didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai proses *opinion mining*.

2.5 *Term Frequency Inverse Document Frequency (TF-IDF)*

TF-IDF merupakan sebuah metode pembobotan yang dilakukan untuk ekstraksi data teks. Tujuan dari *TF-IDF* adalah untuk menemukan jumlah kata yang diketahui (*tf*) setelah dikalikan beberapa banyak tweet dimana suatu kata tersebut muncul (*idf*). Metode *TF-IDF* dilakukan dengan menghitung bobot dengan cara integrase antara *term frequency* (tf) dan *inverse document Frequency* (*idf*) (Kurniawan, 2017). Berikut merupakan rumus untuk menentukan pembobot dengan *TF-IDF*.

$$w_{ij} = tf_{ij} \times (idf + 1) \tag{2.1}$$

$$idf = \log\left[\frac{N}{df_j}\right]$$

dimana w_{ij} adalah bobot dari i pada artikel ke j, N merupakan jumlah seluruh tweet, tf, dan df adalah jumlah tweet j yang mengandung kata i. *TF-IDF* dilakukan agar data dapat di analisis dengan menggunakan Metode Regresi logistik.

2.6 Algoritma Klasifikasi

Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu obyek yang belum diketahui kelasnya (Han dkk., 2012).

2.7 Naïve Bayes Classifier

Teorema Bayes merupakan teorema yang mengacu konsep probabilitas bersyarat (Siang, 2005). Secara umum teorema bayes dapat dinotasikan pada persamaan berikut.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (2.2)$$

Keterangan :

A : sampel data yang label kelasnya tidak diketahui.

B : kelas-kelas hasil klasifikasi.

$P(A|B)$: probabilitas terjadinya A jika B diketahui.

$P(B|A)$: probabilitas terjadinya B jika A diketahui

$P(A)$: probabilitas *prior* A yang mendahului terjadinya B. disebut "prior", karena nilainya bisa diperoleh tanda perlu mempertimbangkan infomasi apapun mengenai B terlebih dahulu, P(A) juga berarti probabilitas ini diperoleh dari data sampel yang telah diketahui berkelas A.

$P(B)$: probabilitas pior B, dan bertindak sebagai *normalizing constant.*

Secara intuitif, terorema *bayes* menggambarakan bahawa perubahan pada "A"dapat diamati apabila "b" terlebih dahulu diamati.

Algoritma *Naïve Bayes Classifier* (NBC) merupakan algoritma yang digunakan untukn mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat (Feldman & Sanger, 2007). Metode *Naïve Bayes Classification* merupakan salah satu metode yang dapat mengklasifikasikan teks. Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki aakurasi yang tinggi. Terdapat dua tahap dalam klasifikasi *tweet*. Tahap pertama adalah pelatihan terhadap tweet yang telah diketahui kategorinya (Falahah & Nur, 2015). Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut

" $a1, a3, \ldots, an$ " dimana $a1$ adalah kata pertama, $a2$ adalah kata kedua dan seterusnya. Sedangakan V adalah himpunan kategori tweet. Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}). Adapun persamaan V_{MAP} adalah sebagai berikut.

$$V_{MAP} = \arg\max_{v_j=V} \mathrm{P}(v_j) \prod_i P(a_i|v_j) \tag{2.3}$$

Nilai $\mathrm{P}(v_j)$ dihitung pada ssaat training, didapat dengan rumus sebagai berikut:

$$P(v_j) = \frac{|\, doc\ j|}{|\, training|} \tag{2.4}$$

dimana $|doc\ j|$ merupakan jumlah *tweet* yang memiliki kategori j dalam training. Sedangkan $|training|$ merupakan jumlah *tweet* dalam yang digunakan untuk training. Untuk setiap probabilitas kata a_i untuk setiap kategori $P(a_i|v_j)$, dihitung pada saat training.

$$P(a_i|v_j) = \frac{n_i+1}{|\, n+kosakata|} \tag{2.5}$$

dimana n_i adalah jumlah kemunculan kata a_i dalam *tweet* yang berkategori v_j dan $|kosakata|$ adalah banyaknya kata dalam training.

2.8 Regresi Logistik

Regresi Logistik adalah salah satu metode pendekatan untuk menemukan hubungan antara variabel prediktor terhadap variabel respon , dengan data pada serangkaian variabel prediktor yang bersifat dikotomus (biner). Pada saat variabel respon adalah variabel kategorik dengan lebih dari dua kategori, maka metode yang digunakan adalah regresi logistik multinomial.

Regresi logistik adalah salah satu regresi dengan fungsi non linear pada parameter sehingga perlu dilakukan transformasi pada $\pi(x)$ dengan transformasi

logit $g(x)$ agar memperoleh fungsi yang linear. Transformasi dilakukan agar dapat ditemukan hubungan antara variabel respon dengan variabel prediktor.

2.9 Regresi Logistik Biner

Model regresi logistic biner digunakan untuk menganalisis hubungan antara satu variabel respond dan beberapa variable predictor jika variable responnya menghasilkan dua kategori bernilai 0 dan 1. Bentuk umum model regresi logistik:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)} \tag{2.6}$$

Fungsi $\pi(x)$ merupakan fungsi non linear sehingga perlu dilakukan transformasi logit untuk memperoleh fungsi yang linear agar dapat dilihat hubungan antara variable respon dengan variabel prediktornya (x). Dengan melakukan transformasi dari logit $\pi(x)$, maka dapat dinyatakan sebagai $g(x)$, yaitu :

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \tag{2.7}$$

dimana:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2.8}$$

Secara umum model probabilitas regresi logistik dengan melibatkan beberapa variabel independen (x) dapat diformulasikan sebagai berikut (Putri, 2018):

$$E(Y|x) = \frac{1}{1+\exp\left(-\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)\right)} \tag{2.9}$$

Sehingga persamaan (4.2) dapat disederhanakan menjadi:

$$E(Y|x) = \frac{1}{1+\exp\left(-\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)\right)}$$

$$= \frac{1}{1+\frac{1}{\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}}$$

$$= \frac{1}{\frac{1+\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}{\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}}$$

$$= \frac{\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)} \quad (2.10)$$

Maka model umum dari regresi logistik dapat dituliskan sebagai berikut (Putri, 2018):

$$\pi(x) = \frac{\exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_k\right)} \quad (2.11)$$

dimana:

$\pi(x)$: peluang untuk variabel prediktor

β : parameter regresi logistik

$x_1, x_2, \ldots, x_p$: variabel prediktor

Untuk menaksir parameter regresinya, maka $\pi(x)$ ditransformasi menggunakan transformasi logit (Putri, 2018).

$$\pi(x) = \frac{\exp\left[g(x)\right]}{1 + \exp\left[g(x)\right]}$$

$$\pi(x) = \left[1 - \pi(x)\right] \exp[g(x)]$$

$$\frac{\pi(x)}{1 - \pi(x)} = \exp[g(x)]$$

$$\ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \ln \exp[g(x)]$$

Sehingga diperoleh transformasi logit sebagai berikut:

$$g(x) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k \quad (2.12)$$

Diketahui probabilitas bersyarat untuk respon dapat dinyatakan sebagai berikut :

$$P_r(Y = 1|x) = \pi(x)$$

dan

$$P_r(Y = 0|x) = 1 - \pi(x)$$

Misal terdapat n buah observasi yang saling bebas. y_i menyatakan variabel respon ke-i, dimana $i = 1,2,3, \dots, n$. Diketahui $\pi(x_i)$ adalah probabilitas untuk $y_i = 1$ dan $1 - \pi(x_i)$ adalah probabilitas untuk $y_i = 0$. Maka fungsi kepadatan peluang dari y_i adalah :

$$f(y_i) = [\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i} \quad (2.13)$$

Fungsi *likelihood* diperoleh dengan mengalikan fungsi-fungsi kepadatan peluang (pdf) dari y_i dimana untuk setiap y_i diasumsikan saling bebas, maka:

$$\begin{aligned} L(\beta) &= f(y_1, y_2, \dots, y_n) \\ &= f(y_1) \cdot f(y_2) \cdot \dots \cdot f(y_n) \\ &= \prod_{i=1}^{n}[\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i} \end{aligned} \quad (2.14)$$

Untuk mendapatkan nilai $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ yang memaksimumkan fungsi *likelihood* bentuk logaritma natural dari fungsi *likelihood* (Putri, 2018).

$$\begin{aligned} l(\beta) &= \ln\big(L(\beta)\big) \\ &= \ln\left(\prod_{i=1}^{n}[\pi(x_i)]^{y_i}[1 - \pi(x_i)]^{1-y_i}\right) \\ &= \sum_{i=1}^{n}\big\{y_i \ln \pi(x_i) + (1 - y_i)\ln\big(1 - \pi(x_i)\big)\big\} \\ &= \sum_{i=1}^{n}\left\{y_i \ln\left(\frac{\exp(g(x_i))}{1+\exp(g(x_i))}\right) + (1 - y_i)\ln\left(1 - \frac{\exp\left(g(x_i)\right)}{1+\exp\left(g(x_i)\right)}\right)\right\} \\ &= \sum_{i=1}^{n}\big\{y_i\big[\ln\exp\big(g(x_i)\big) - \ln\big(1 + \exp\big(g(x_i)\big)\big)\big] \\ &\qquad + (1 - y_i)[(\ln 1) - \ln(1 + \exp g(x_i))]\big\} \\ &= \sum_{i=1}^{n}\left\{y_i\big[\ln\exp\big(g(x_i)\big) - \ln\big(1 + \exp\big(g(x_i)\big)\big)\big]\right. \\ &\qquad \left. - (1 - y_i)\ln\left(1 + \exp\left(g(x_i)\right)\right)\right\} \end{aligned}$$

$$= \sum_{i=1}^{n} \left\{ y_i \ln\left(\exp\left(g(x_i)\right)\right) - y_i \ln\left(1 + \exp\left(g(x_i)\right)\right) - \ln\left(1 + \exp\left(g(x_i)\right)\right) + y_i \ln\left(1 + \exp\left(g(x_i)\right)\right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i\left(g(x_i)\right) - \ln\left(1 + \exp\left(g(x_i)\right)\right) \right\}$$

Maka, fungsi *log-likelihood* dapat dituliskan menjadi:

$$l(\beta) = \sum_{i=1}^{n} \left\{ y_i\left(g(x_i)\right) - \ln\left(1 + \exp\left(g(x_i)\right)\right) \right\} \quad (2.15)$$

Selanjutnya memaksimumkan persamaan (2.15) dengan menurunkan fungsi *log-likelihood* terhadap β_k yang disamakan dengan nol (Putri, 2018).

$$\frac{\partial l(\beta)}{\partial \beta_0} = 0$$

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \left\{ y_i - \exp\left(g(x_i)\right) \left(\frac{1}{1 + \exp\left(g(x_i)\right)} \right) \right\} = 0$$

$$= \sum_{i=1}^{n} \left\{ y_i - \left(\frac{\exp\left(g(x_i)\right)}{1 + \exp\left(g(x_i)\right)} \right) \right\} = 0$$

$$\frac{\partial l(\beta)}{\partial k} = 0$$

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} \left\{ y_i x_{ik} - x_{ik} \exp\left(g(x_i)\right) \left(\frac{1}{1 + \exp\left(g(x_i)\right)} \right) \right\} = 0$$

$$= \sum_{i=1}^{n} \left\{ y_i x_{ik} - x_{ik} \left(\frac{\exp\left(g(x_i)\right)}{1 + \exp\left(g(x_i)\right)} \right) \right\} = 0$$

$$= \sum_{i=1}^{n} \left\{ x_{ik} \left[y_i - \left(\frac{\exp\left(g(x_i)\right)}{1 + \exp\left(g(x_i)\right)} \right) \right] \right\} = 0 \quad (2.16)$$

dimana $k = 1,2, \dots, p$

Metode *Maksimum Likelihood Estimation* diperoleh hasil turunan β disamakan dengan nol membentuk suatu sistem persamaan yang tidak linear dan tidak dapat diselesaikan secara analitik tetapi didekati secara iteratif dengan menggunakan metode Newton Raphson. Dasar dari metode ini adalah pendekatan deret Taylor yang dituliskan sebagai berikut:

$$f(x_{t+1}) = f(x_t) + f'(x_t)(x_{t+1} - x_t) + \frac{f''(x_t)}{2!}(x_{t+1} - x_t)^2 + \cdots + \frac{f''(x_t)}{n!}(x_{t+1} - x_t)^n$$

Metode Newton Raphson membutuhkan turunan pertama dan ke dua dari fungsi *log-likelihood* karena metode Newton Raphson menggunakan pendekatan deret Taylor orde ke dua. Deret Taylor orde ke dua dapat ditunjukkan dalam bentuk:

$$f(x_{t+1}) = f(x_t) + f'(x_t)(x_{t+1} - x_t) + \frac{f''(x_t)}{2!}(x_{t+1} - x_t)^2$$

Misalkan $f(a)$ adalah deret Taylor orde ke dua dan diketahui $a = x_{t+1}$ dan $b = x_t$, maka diperoleh:

$$f(a) = f(b) + f'(b)(a - b) + \frac{f''(b)}{2!}(a - b)^2$$

Agar nilai $f(a)$ maksimum, maka harus memenuhi $\frac{\partial f(a)}{\partial a} = 0$

$$\frac{\partial f(a)}{\partial a} = 0$$

$$0 = 0 + f'(b) + \frac{f''(b)}{2!}2(a - b)$$

$$f''(b)(a - b) = -f'(b)$$

$$[f''(b)]^{-1} \cdot f''(b)(a - b) = -[f''(b)]^{-1} \cdot f'(b)$$

$$I(a - b) = -[f''(b)]^{-1} \cdot f'(b)$$

$$a = b - [f''(b)]^{-1} \cdot f'(b)$$

Sehingga diperoleh rumus penaksiran parameter β dapat ditunjukkan dalam bentuk:

$$\begin{bmatrix} \hat{\beta}_{0(t+1)} \\ \hat{\beta}_{1(t+1)} \\ \vdots \\ \hat{\beta}_{p(t+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{0(t)} \\ \hat{\beta}_{1(t)} \\ \vdots \\ \hat{\beta}_{p(t)} \end{bmatrix} - [f''(b)]^{-1} \cdot f'(b)$$

Persamaan (2.16) merupakan turunan pertama dari fungsi *log-likelihood*, selanjutnya untuk turunan ke dua (Putri, 2018):

$$\frac{\partial^2 l(\beta)}{\partial \beta_0^2} = 0$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{[\exp(g(x_i))(1+\exp(g(x_i)))] - [\exp(g(x_i)) \cdot \exp(g(x_i))]}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{\exp(g(x_i)) + \exp(2(g(x_i))) - \exp(2(g(x_i)))}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{\exp(g(x_i))}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_k^2} = 0$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{(x_{ik}) x_{ik} \exp(g(x_i))[1+\exp(g(x_i))] - [x_{ik} \cdot \exp(g(x_i)) \cdot x_{ik} \exp(g(x_i))]}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{{x_{ik}}^2 \exp(g(x_i)) + {x_{ik}}^2 \exp(2(g(x_i))) - x_{ik} \exp(2(g(x_i)))}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{{x_{ik}}^2 \exp(g(x_i))}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_k} = 0$$

$$0 = -\sum_{i=1}^{n} \left\{ \left(\frac{[x_{ik} \exp(g(x_i))(1+\exp(g(x_i)))] - [\exp(g(x_i)) \cdot x_{ik} \exp(g(x_i))]}{[1+\exp(g(x_i))]^2} \right) \right\}$$

$$0 = -\sum_{i=1}^{n}\left\{\left(\frac{x_{ik}\exp(g(x_i))+x_{ik}\exp\left(2(g(x_i))\right)-x_{ik}\exp\left(2(g(x_i))\right)}{\left[1+\exp(g(x_i))\right]^2}\right)\right\}$$

$$0 = -\sum_{i=1}^{n}\left\{\left(\frac{x_{ik}\exp(g(x_i))}{\left[1+\exp(g(x_i))\right]^2}\right)\right\}$$

$$\frac{\partial^2 l(\beta)}{\partial\beta_q\partial\beta_k} = 0$$

$$0 = -\sum_{i=1}^{n}\left\{\left(\frac{x_{iq}\cdot x_{ik}\exp(g(x_i))[1+\exp(g(x_i))]-[x_{iq}\exp(g(x_i))\cdot x_{ik}\exp(g(x_i))]}{\left[1+\exp(g(x_i))\right]^2}\right)\right\}$$

$$0 = \sum_{i=1}^{n}\left\{\left(\frac{xi_q\cdot x_{ik}\exp(g(x_i))+x_q\cdot x_{ik}\exp\left(2(g(x_i))\right)-x_{iq}\cdot x_{ik}\exp\left(2(g(x_i))\right)}{\left[1+\exp(g(x_i))\right]^2}\right)\right\}$$

$$0 = -\sum_{i=1}^{n}\left\{\left(\frac{xi_q\cdot x_{ik}\exp(g(x_i))}{\left[1+\exp(g(x_i))\right]^2}\right)\right\} \quad (2.17)$$

dimana $q = k = 1, 2, \dots, p$

Proses iterasi Newton-Raphson akan berhenti jika terpenuhi kondisi konvergen, yaitu selisih $\left\|\beta^{(t+1)} - \beta^{(t)}\right\| \leq \varepsilon$, dimana ε adalah bilangan positif yang ditoleransi.

2.9.1 Pengujian Parameter

Uji signifikansi parameter dari variabel prediktor dilakukan untuk mengetahui apakah taksiran parameter yang diperoleh berpengaruh secara signifikan terhadap model atau tidak, dan seberapa besar pengaruh masing-masing parameter tersebut terhadap model. Uji signifikansi terdiri dari dua tahap yaitu uji signifikansi parameter model secara individu dan uji signifikansi parameter model secara bersama.

2.9.1.1 Uji Serentak

Pengujian secara serentak dilakukan untuk mengetahui peranan setiap variabel prediktor dalam model secara keseluruhan (serentak).

Hipotesis:

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (tidak ada pengaruh antara variabel prediktor dengan variabel respon)

H_1 : Paling sedikit ada satu $\beta_j \neq 0$, dengan j = 1,2, … , p (terdapat paling sedikit satu variabel prediktor yang berpengaruh terhadap variabel respon)

Statistik uji yang digunakan adalah statistik uji rasio *Likelihood* yang disimbolkan *G.* Statistik uji *G* ini berdistribusi *chi-square* dengan derajat bebas p dimana p merupakan banyaknya variabel prediktor.

Statistik uji (Hosmer & Lemeshow, 2000):

$$G^2 = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}}\right] \quad (2.18)$$

dengan,

$n_0 = \sum_{i=1}^{n}(1 - y_i)$

$n_1 = \sum_{i=1}^{n} y_i$

$n = n_0 + n_1$

dimana:

n_0 : banyaknya nilai observasi $y = 0$

n_1 : banyaknya nilai observasi $y = 1$

n : banyaknya observasi

Kriteria keputusan pengujian dilakukan dengan membandingkan nilai statistik uji G^2 dengan nilai X^2 tabel *chi-square* dengan derajat bebas pada taraf signifikan α. H_0 ditolak jika nilai statistik uji $G > X^2_{(p,\alpha)}$ atau dengan melihat nilai *p-value*. Jika nilai *p-value*$< \alpha$, maka H_0 ditolak.

2.9.1.2 Uji Kecocokan Model

Keefektifan dari suatu model dalam menjelaskan variabel respon dapat diketahui berdasarkan estimasi model regresi logistik yang telah diperoleh. Uji kecocokan model digunakan untuk mengevaluasi cocok tidaknya model dengan data dan nilai observasi yang diperoleh sama atau mendekati dengan yang diharapkan dalam model. Hal ini dapat disebut sebagai *Goodness of fit* (kesesuian model). *Goodness of fit* dihitung berdasarkan $\hat{\pi}$ yang tergantung pada susunan variabel-variabel prediktor dalam model, bukan dari jumlah variabel prediktor (Hosmer & Lemeshow, 2000).

Hipotesis:

H_0 : Model sesuai (tidak ada perbedaan antara prediksi dan hasil observasi)

H_1 : Model tidak sesuai (ada perbedaan antara prediksi dan hasil observasi)

Statistik uji:

$$\hat{C} = \sum_{i=1}^{g} \frac{(o_p - n_p{}'\bar{\pi}_p)}{n_p{}'\bar{\pi}_p(1-\pi_p)} \tag{2.19}$$

dimana:

g : jumlah grup

$n_p{}'$: jumlah subjek pada grup ke- p

o_p : jumlah nilai variabel respon pada grup ke- p

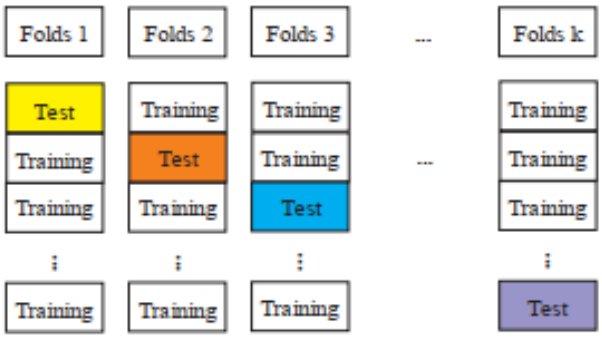
$\bar{\pi}_p$: rata-rata taksiran probabilitas dimana m_j adalah banyaknya subjek pada c_p kombinasi variabel prediktor.

Jika H_0 diterima, maka distribusi statistik uji mengikuti distribusi *chi-square* dengan derajat bebas $(g-2)$. Daerah penolakan H_0 adalah $\hat{C} > X^2{}_{(g-2)}$ atau dengan menggunakan nilai *deviance* dengan nilai *p-value* $< \alpha$. Dapat disimpulkan jika H_0 diterima bahwa model yang diperoleh telah sesuai dan

sebaliknya jika H_0 ditolak maka disimpulkan bahwa model yang diperoleh belum sesuai.

2.10 *K-Fold Cross Validation*

K-fold cross validation adalah salah satu metode yang digunakan untuk mempartisi data menjadi data *training* dan data *testing*. Metode ini banyak digunakan peneliti karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. *K-fold cross validation* secara berulang-ulang membagi data menjadi data training dan data testing, dimana setiap data mendapat kesempatan menjadi data testing (Gokgoz & Subasi, 2015). K merupakan besar angka partisi data yang digunakan untuk pembagian training-testing. Berikut merupakan ilustrasi pembagian data menggunakan *k-fold cross validation.*



Gambar 2.4 Ilustrasi *k-fold cross validation*.

2.11 Ketepatan Klasifikasi

Pengukuran ketetapan klasifikasi dilakukan untuk melihat performa klasifikasi yang telah dilakukan. Dalam mengukur ketetapan klasifikasi, perlu diketahui jumlah pada setiap kelas prediksi dan kelas actual yang terdiri dari *TP* (*True Positive*) yaitu jumlah tweet bersentimen positif yang tepat terprediksi dalam kelas positif, *TN* (*True Negative*) yaitu *tweet* bersentimen positif yang terprediksi dalam kelas negative (Kurniawan, 2017). Berikut merupakan *confusion matrix* yang memuat keempat nilai tersebut.

Kelas Aktual	Kelas Prediksi	
	Positif	Negatif
Positif	*TP*	*FN*
Negatif	*FP*	*TN*

Gambar 2.5 *Confusion matrix*

Pengukuran yang sering digunakan untuk digunakan untuk menghitung ketepatan klasifikasi adalah akurasi, *specificity*, dan *sensitivity/recall* (Hotho dkk., 2005). Akurasi merupakan presentase dokumen yang teridentifikasi secara tepat dari total dokumen dalam proses klasifikasi. Akurasi digunakan untuk menghitung ketepatan kalsifikasi sebuah dokumen yang mempunyai data yang *balanced* pada tiap kategorinya. Berikut merupakan rumus dalam menghitung akurasi, *sepecivity* dan *sensitivity/recall.*

$$Akurasi = \frac{TN+TP}{TN+TP+FN+FP} \tag{2.20}$$

$$Sensitivity/Recall = \frac{TP}{TP+FN} \tag{2.21}$$

$$Specivity = \frac{TN}{TN+FP} \tag{2.22}$$

Untuk *imbalanced-data* maka ketepatan klasifikasi akan difokuskan pada nilai *sensitivity/Recall* dan *Specitivity* untuk melihat akurasi prediksi setiap *class*.

2.12 Visualisasi

Visualisasi adalah suatu rekayasa dalam pembuatan gambar, diagram atau animasi untuk penampilan suatu informasi. Secara umum, Visualisasi dalam bentuk gambar baik yang bersifat abstrak maupun nyata telah dikenal sejak awal peradaban manusia (Olomouc, 2008).

2.12.1 World Cloud

World Cloud merupakan salah satu metode visualisasi dokumen teks yang sering digunakan. *Word cloud* merupakan representasi grafis dari sebuah dokumen yang dilakukan dengan plotting kata-kata yang sering muncul ditunjukkan melalui

ukuran huruf kata tersebut. Semakin besar ukuran kata menunjukkan semakin besar frekuensi kata tersebut muncul dalam dokumen (Kurniawan, 2017). Berikut merupakan contoh dari visualisasi dokumen teks dengan *world cloud*

Gambar 2.6 Visualisasi *World Cloud*

2.12.2 Visualisasi Pemetaan (Geospasial)

Informasi lokasi biasanya digunakan untuk mendapatkan wawasan tentang lokasi terkemuka yang mendiskusikan suatu kasus. Peta adalah pilihan yang jelas untuk memvisualisasikan informasi lokasi. Di bagian ini, kita akan membahas bagaimana peta dapat digunakan untuk meringkas informasi lokasi secara efektif dan membantu dalam analisis Tweet.

Upaya pertama untuk membuat peta yang mengidentifikasi lokasi Tweet adalah dengan menyorot masing-masing lokasi Tweet. Setiap Tweet diidentifikasi oleh sebuah titik di peta, dan titik-titik tersebut disebut sebagai penanda. Biasanya, bentuk, warna, dan gaya penanda dapat disesuaikan agar sesuai dengan persyaratan aplikasi. Peta ditampilkan sebagai kumpulan gambar, yang disebut ubin. Contoh dari pendekatan "titik di peta" disajikan pada Gambar 2.6 yang menunjukkan lokasi di Google Maps. Sentimen tweet juga digunakan dalam memvisualisasikan data, Jika tweet positif maka divisualisasikan dengan titik hijau, jika tweet negatif maka divisualisasikan menggunakan titik merah dan titik abu-abu untuk tweet yang secara sentimental netral (Kotrika, 2016).

Gambar 2.7 Visualisasi Geospasial