

**ANALISIS SENTIMEN MASYARAKAT TERHADAP E-COMMERCE
PADA MEDIA SOSIAL MENGGUNAKAN METODE
NAÏVE BAYES CLASSIFIER (NBC) DENGAN
SELEKSI FITUR INFORMATION GAIN (IG)**

REVISI USULAN SKRIPSI



Oleh :

ABDAN SYAKURO
13650129

**JURUSAN TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2017**

**ANALISIS SENTIMEN MASYARAKAT TERHADAP E-COMMERCE
PADA MEDIA SOSIAL MENGGUNAKAN METODE
NAÏVE BAYES CLASSIFIER (NBC) DENGAN
SELEKSI FITUR INFORMATION GAIN (IG)**

REVISI USULAN SKRIPSI



Oleh :

ABDAN SYAKURO
13650129

Telah disetujui pada tanggal : 2017

Pembimbing I

Yang Mengajukan

(Fachrul Kurniawan, ST., M. MT)
NIP. 19771020 200901 1 001

(Abdan Syakuro)
NIM. 13650129

LEMBAR PENGESAHAN

REVISI USULAN SKRIPSI

ANALISIS SENTIMEN MASYARAKAT TERHADAP E-COMMERCE PADA MEDIA SOSIAL MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER (NBC) DENGAN SELEKSI FITUR INFORMATION GAIN (IG)



Oleh :

ABDAN SYAKURO
13650129

Pada Tanggal : 2017

Telah diperiksa, disetujui, dan disahkan oleh :

Penguji I	: (<u>Dr. M. Faisal, MT</u>)	
	19740510 200501 1 007	()
Penguji II	: (<u>Fresy Nugroho, MT</u>)	
	19710722 201101 1 001	()
Pembimbing I	: (<u>Fachrul Kurniawan, M.MT</u>)	
	19771020 200901 1 001	()
Pembimbing II	: (<u>A'la Syauqi, M.Kom</u>)	
	19771201 200801 1 007	()

DAFTAR ISI

BAB I.....	1
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	3
1.5 Batasan Masalah	4
BAB II.....	4
TINJAUAN PUSTAKA	4
2.1 Penelitian Terdahulu	4
2.2 <i>Data Mining</i>	5
2.3 <i>Text Mining</i>	6
2.4 Analisis Sentimen	6
2.5 Media Sosial.....	7
2.6 <i>Text Preprocessing</i>	7
2.7 Seleksi Fitur	8
2.8 <i>Information Gain (IG)</i>	8
2.9 <i>Naïve Bayes Classifier (NB)</i>	9
2.10 Evaluasi.....	11
BAB III	13
3.1 Deskripsi Umum	13
3.2 <i>Input</i>	13
3.3 <i>Dataset</i>	14
3.4 <i>Preprocessing</i>	15
3.4.1 <i>Case Folding</i>	16
3.4.2 <i>Cleansing</i>	16
3.4.3 <i>Convert Emoticon</i>	17
3.4.4 <i>Convert Negation</i>	18
3.4.5 <i>Tokenizing</i>	19
3.4.6 <i>Normalization</i>	20

3.4.7	<i>Stopword Removal</i>	22
3.4.8	<i>Stemming</i>	23
3.4.9	Seleksi Fitur <i>Information Gain</i>	25
3.5	<i>Naïve Bayes Classification</i>	27
3.6	Perancangan Desain Sistem	33
3.7	Hasil Analisis Sistem	34
DAFTAR PUSTAKA		35

BAB I

PENDAHULUAN

1.1 Latar Belakang

Seiring berkembangnya teknologi internet di Indonesia, memunculkan banyak situs jual beli *online* atau yang lebih populer disebut *e-commerce*. Saat ini *e-commerce* merupakan tempat jual beli *online* yang semakin diminati oleh masyarakat kota. Hal ini dikarenakan kemudahan transaksi tanpa harus datang ke toko fisik, dan setiap tahunnya jumlah pengguna *e-commerce* semakin meningkat. Menurut data yang dirilis biro riset Frost & Sullivan, bersama China, Indonesia menjadi negara dengan pertumbuhan pasar *e-commerce* terbesar di dunia dengan rata – rata pertumbuhan 17 persen setiap tahun. (tekno.liputan6.com, 2015)

Dari peningkatan jumlah pengguna *e-commerce* tersebut, tentunya kejahatan atau penipuan dalam dunia *online* juga semakin meningkat. Dalam riset yang dilakukan oleh Kaspersky Lab dan B2B *International*, terungkap bahwa 48% konsumen menjadi target aksi penipuan yang dirancang untuk menipu dan mengelabui mereka sehingga mengungkapkan informasi sensitif dan data keuangan untuk tindak kriminal. Hal yang mengkhawatirkan, dari 26 negara yang disurvei, Indonesia menempati posisi tertinggi sebesar 26% konsumen telah kehilangan uang mereka sebagai akibat menjadi target aksi penipuan *online* (inet.detik.com, 2016). Dari banyaknya *e-commerce* yang ada di Indonesia juga membuat persaingan dalam hal pelayanan, karena pelayanan sangat berpengaruh terhadap kepuasan konsumen, sehingga penilaian masyarakat terhadap suatu produk dapat dijadikan analisa terhadap pasar *online*.

Untuk mengatasi masalah tersebut, opini masyarakat terhadap suatu produk dapat membantu masyarakat lain supaya lebih berhati – hati dalam transaksi *online*. Penulis merasa perlu untuk melakukan penelitian ini yaitu dengan membuat sebuah sistem penganalisa opini atau biasa disebut sentimen analisis masyarakat sehingga bisa mengetahui dan membantu memberikan informasi mengenai analisa sentimen *e-commerce* masyarakat.

Adapun *e-commerce* yang menjadi objek penelitian ini adalah Lazada, Bukalapak, dan Tokopedia. Dilansir dari artikel yang ditulis oleh media *online*

tekno.liputan6.com, ketiga *e-commerce* tersebut termasuk lima situs *e-commerce* terbaik di Indonesia yang paling sering dikunjungi konsumen. Data tersebut diperoleh dari situs Alexa dengan peringkat pertama adalah Bukalapak, yang kedua adalah Lazada dan yang ketiga adalah Tokopedia.

Opini masyarakat dapat diperoleh dari berbagai media cetak maupun elektronik. Masyarakat kota saat ini lebih sering menggunakan media sosial dalam mengomentari suatu masalah termasuk suatu produk. Salah satu media sosial yang digemari masyarakat Indonesia saat ini adalah *Twitter*.

Orang Indonesia dikenal sangat aktif di media sosial. *Country Business Head Twitter* Indonesia Roy Simangunsong mengatakan, jumlah cuitan orang Indonesia selama Januari hingga Desember 2016 mencapai 4,1 miliar *tweet*. Meski tak menyebut jumlah pengguna di Indonesia, Roy mengatakan bahwa jumlah pengguna aktif *Twitter* Indonesia mencapai 77 persen dari seluruh pengguna di dunia (tekno.liputan6.com, 2016). Tentu saja, informasi yang terkandung dalam *tweet* ini sangat berharga sebagai alat penentu kebijakan dan ini bisa dilakukan dengan *Text Mining*.

Text Mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi dokumen dimana *Text Mining* merupakan variasi dari *Data Mining* yang berusaha menemukan pola menarik dari sekumpulan data tekstual yang berjumlah besar (Feldman & Sanger, 2007). (Kurniawan, Effendi, & Sitompul, 2012)

Analisis Sentimen atau *Opinion Mining* adalah studi komputasional dari opini – opini orang, sentimen dan emosi melalui entitas atau atribut yang dimiliki yang diekspresikan dalam bentuk teks (Liu, 2012). (Aditya, Hani'ah, Fitrawan, Arifin, & Purwitasari, 2016). Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam kalimat atau dokumen untuk mengetahui pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif, atau netral (Pang & Lee, Opinion mining and sentiment analysis, 2008).

Masalah sentimen analisis sangat penting untuk diperhatikan oleh seorang muslim karena Allah berfirman dalam Al-Qur'an ayat 6 berfirman :

"Hai orang-orang yang beriman, jika datang kepadamu orang fasik, membawa suatu berita, maka periksalah dengan teliti, agar kamu tidak menimpakan suatu musibah, kepada suatu kaum, tanpa mengetahui keadaannya, yang menyebabkan kamu menyesal atas perbuatanmu itu."

Ayat tersebut menerangkan kepada kita bahwa jika kita mendapatkan opini masyarakat yang belum jelas, kita harus melakukan tabayyun dahulu. Karena jika kita tidak tabayyun dapat mengakibatkan dampak buruk bagi kita sendiri maupun orang lain. Berangkat dari ayat tersebut, maka penulis membangun sebuah alat yang salah satu fungsinya adalah tabayyun dengan mengumpulkan berbagai opini dari masyarakat lalu menganalisisnya sehingga mendapatkan hasil yang berguna bagi masyarakat.

Berdasarkan penelitian (Chandani, Wahono, & Purwanto, 2015) yang meneliti perbandingan akurasi antara *information gain*, *chi square*, *forward selection*, *backward selection* didapatkan *information gain* merupakan yang terbaik. Sehingga pada penelitian ini penulis akan membuat aplikasi Analisis Sentimen menggunakan metode *Naïve Bayes* untuk klasifikasi dengan seleksi *fitur information gain*.

1.2 Rumusan Masalah

- 1) Bagaimana membangun aplikasi *sentimen analisis* dari media sosial menggunakan metode *Naïve Bayes* dengan seleksi *fitur Information Gain* terhadap *e-commerce*?
- 2) Bagaimana membangun analisa data berbasis sentimen analisis dari media sosial terhadap *e-commerce*?

1.3 Tujuan Penelitian

- 1) Membangun aplikasi klasifikasi sentimen masyarakat terhadap *e-commerce* menggunakan metode *Naïve Bayes* dengan seleksi *fitur Information Gain*.
- 2) Membangun analisa data berbasis sentimen analisis untuk membantu menambah informasi masyarakat dalam memilih belanja di *e-commerce*.

1.4 Manfaat Penelitian

- 1) Memberikan khazanah keilmuan dan menambah informasi khususnya pertimbangan dalam permasalahan – permasalahan *e-commerce*.
- 2) Memberikan informasi sentimen positif atau negatif kepada pemilik produk agar bisa menganalisa produknya.

1.5 Batasan Masalah

- 1) Data yang dianalisis adalah data dari media sosial berbahasa Indonesia.
- 2) Metode yang digunakan untuk klasifikasi adalah *Naïve Bayes* dengan seleksi fitur *Information Gain*.
- 3) Data yang dijadikan data *testing* hanya 8 hari ke belakang dari hari proses *testing*.
- 4) Jumlah data *training* yang digunakan 3000 *tweet*.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian mengenai klasifikasi sentimen telah dilakukan oleh Bo Pang (2002). Pada jurnalnya, Bo Pang melakukan klasifikasi sentimen terhadap *review* film dengan menggunakan berbagai teknik pembelajaran mesin. Teknik pembelajaran mesin yang digunakan yaitu *Naïve Bayes*, *Maximum Entropy*, dan *Support Vector Machines (SVM)*. Pada penelitian itu juga digunakan beberapa pendekatan untuk melakukan ekstraksi *fitur*, yaitu *unigram*, *unigram + bigram*, *unigram + Part of Speech (POS)*, *adjective*, dan *unigram + posisi*. Hasil dari eksperimen yang dilakukan di penelitian ini menemukan bahwa *SVM* menjadi metode terbaik ketika dikombinasikan dengan *unigram* dengan akurasi 82.9% (Pang & Lee, Thumbs up? Sentiment Classification using Machine Learning, 2002).

Pada penelitian yang dilakukan oleh Rozi, I. F., Pramono, S. H., & Dahlan, E. A. (2012) dikembangkan sistem *opinion mining* untuk menganalisis opini publik pada perguruan tinggi. Pada *subproses document subjectivity* dan *target detection* digunakan *Part-of-Speech (POS) Tagging* menggunakan *Hidden Markov Model (HMM)*. Pada hasil proses *POS Tagging* kemudian diterapkan *rule* untuk mengetahui apakah suatu dokumen termasuk opini atau bukan, serta untuk mengetahui bagian kalimat mana yang merupakan objek yang menjadi target opini. Dokumen yang dikenali sebagai opini selanjutnya diklasifikasikan ke dalam opini negatif dan positif (*subproses opinion orientation*) menggunakan *Naïve Bayes Classifier (NBC)*. Dari pengujian didapatkan nilai *precision* dan *recall* untuk *subproses document subjectivity* adalah 0.99 dan 0.88, untuk *subproses target detection* adalah 0.92 dan 0.93, serta untuk *subproses opinion orientation* adalah 0.95 dan 0.94. (Rozi, Pranomo, & Dahlan, 2012)

Penelitian tentang klasifikasi juga dilakukan oleh Rodiyansyah, S. F. dan Winarko Edi dengan melakukan teknik *Data Mining* yang digunakan untuk visualisasi kemacetan lalu lintas di sebuah kota. Pada penelitiannya ini metode yang dipakai adalah *Naïve Bayes* dengan mengkombinasikan pengetahuan

sebelumnya dengan pengetahuan baru. Dari hasil uji coba, aplikasi menunjukkan bahwa nilai akurasi terkecil 78% dihasilkan pada pengujian dengan sampel sebanyak 100 dan menghasilkan nilai akurasi tinggi 91,60% pada pengujian dengan sampel sebanyak 13106. Hasil pengujian dengan perangkat lunak *Rapid Miner 5.1* diperoleh nilai akurasi terkecil 72% dengan sampel sebanyak 100 dan nilai akurasi tertinggi 93,58% dengan sampel 13106 untuk metode *Naive Bayesian Classification*. Sedangkan untuk metode *Support Vector Machine* diperoleh nilai akurasi terkecil 92% dengan sampel sebanyak 100 dan nilai akurasi tertinggi 99,11% dengan sampel sebanyak 13106. (Rodiyansyah & Winarko, 2013)

Penelitian lainnya pada klasifikasi sentimen *review* film dilakukan oleh Chandani, V., Wahono, R. S., & Purwanto (2015) dengan mengkomparasi metode klasifikasi *Machine Learning* seperti *Naïve Bayes (NB)*, *Support Vector Machine (SVM)*, dan *Artificial Neural Network (ANN)* dan Seleksi Fitur seperti *Information Gain*, *Chi Square*, *Forward Selection* dan *Backward Elimination*. Hasil komparasi metode *SVM* mendapatkan hasil yang terbaik dengan akurasi 81.10% dan *AUC* 0.904. Hasil dari komparasi Seleksi Fitur *Information Gain* mendapatkan hasil yang paling baik dengan rata – rata akurasi 84.57% dan rata – rata *AUC* 0.899. Hasil integrasi metode klasifikasi terbaik dan metode Seleksi Fitur terbaik menghasilkan akurasi 81.50% dan *AUC* 0.929. Hasil ini mengalami kenaikan jika dibandingkan hasil eksperimen yang menggunakan *SVM* tanpa Seleksi Fitur. Hasil dari pengujian metode Seleksi Fitur terbaik untuk setiap metode klasifikasi adalah *Information Gain* mendapatkan hasil terbaik untuk digunakan pada metode *NB*, *SVM* dan *ANN* (Chandani, Wahono, & Purwanto, 2015).

2.2 Data Mining

Data Mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran computer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis.

Data Mining merupakan proses *iteratif* dan interaktif untuk mengemukakan pola atau model baru sah (sempurna), bermanfaat dan dapat dimengerti dalam suatu *database* yang sangat besar (*massive database*).

Data Mining berisi pencarian tren atau pola yang diinginkan dalam *database* besar untuk membantu pengambilan keputusan di waktu yang akan datang. Pola – pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisa data berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin saja menggunakan perangkat pendukung keputusan yang lainnya (Hermawati, 2013).

2.3 *Text Mining*

Text Mining adalah proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata – kata yang dapat mewakili isi dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen tersebut. (Aditya B. R., 2015)

Text Mining juga dikenal sebagai *Data Mining Text* atau penemuan pengetahuan dari *database* tekstual. Sesuai dengan buku *The Text Mining Handbook*, *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*. Tujuan dari *Text Mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan dalam *Text Mining* adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal *semi* terstruktur. Adapun tugas khusus dari *Text Mining* antara lain yaitu pengkategorisasian teks dan pengelompokkan teks. (Nurhuda, Sihwi, & Doewes, 2013)

2.4 Analisis Sentimen

Menurut Medhat et al (2014, 1093) Analisis sentimen adalah suatu bidang yang berlangsung dalam penelitian berbasiskan teks. Analisis sentimen atau *opinion mining* adalah kajian tentang cara untuk memecahkan masalah dari opini masyarakat, sikap dan emosi suatu entitas, dimana entitas tersebut dapat mewakili individu. (Wati, 2016)

Analisis Sentimen atau *opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah

masalah atau objek oleh seseorang, apakah cenderung beropini negatif atau positif. (Rozi, Pranomo, & Dahlan, 2012)

2.5 Media Sosial

Pada dasarnya media sosial merupakan perkembangan mutakhir dari teknologi – teknologi web baru berbasis internet, yang memudahkan semua orang untuk dapat berkomunikasi, berpartisipasi, saling berbagai dan membentuk sebuah jaringan secara *online*, sehingga dapat menyebarluaskan konten mereka sendiri. *Post* di blog, *tweet*, atau video *youtube* dapat direproduksi dan dapat dilihat secara langsung oleh jutaan orang secara gratis.

Media sosial mempunyai banyak bentuk, diantaranya yang paling populer yaitu *microblogging* (*twitter*), facebook, dan blog. *Twitter* adalah suatu situs web yang merupakan layanan dari *microblog*, yaitu suatu bentuk blog yang membatasi ukuran setiap *post*-nya, yang memberikan fasilitas bagi pengguna untuk dapat menuliskan pesan dalam *twitter update* hanya berisi 140 karakter. *Twitter* merupakan salah satu jejaring sosial yang paling mudah digunakan, karena hanya memerlukan waktu yang singkat tetapi informasi yang disampaikan dapat langsung menyebar secara luas. (Setyani, 2013)

2.6 Text Preprocessing

Preprocessing merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, *preprocessing* data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. *Preprocessing* sangat penting dalam pembuatan *sentimen analisis*, terutama untuk media sosial yang sebagian besar berisi kata – kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar. Ada tiga model *preprocessing* untuk kalimat atau teks dengan *noise* yang besar (A Clark, 2003). Tiga model tersebut adalah :

1. *Orthographic Model*. Model ini dipergunakan untuk memperbaiki kata atau kalimat yang memiliki kesalahan dari segi bentuk kata atau kalimat. Contoh kesalahan yang diperbaiki dengan *Orthographic model* adalah huruf kapital di tengah kata.
2. *Error Model*. Model ini dipergunakan untuk memperbaiki kesalahan dari segi kesalahan eja atau kesalahan penulisan. Ada dua jenis

kesalahan yang dikoreksi dengan model ini yaitu kesalahan penulisan dan kesalahan eja. Kesalahan penulisan mengacu pada kesalahan pengetikan sedangkan kesalahan eja muncul ketika penulis tidak tahu ejaannya benar atau salah.

3. *White Space Model*. Model ke tiga ini mengacu pada pengoreksian tanda baca. Contoh kesalahan untuk model ini adalah tidak menggunakan tanda titik ‘.’ di akhir kalimat. Namun, model ini tidak terlalu signifikan, terutama ketika berhadapan dengan media sosial yang jarang mengindahkan tanda baca. (Mujilahwati, 2016)

2.7 Seleksi Fitur

Seleksi *fitur* adalah salah satu teknik terpenting dan sering digunakan dalam pre-processing. Teknik ini mengurangi jumlah *fitur* yang terlibat dalam menentukan suatu nilai kelas target, mengurangi *fitur irrelevant*, berlebihan dan data yang menyebabkan salah pengertian terhadap kelas target yang membuat efek segera bagi aplikasi. Tujuan utama dari seleksi *fitur* ialah memilih *fitur* terbaik dari suatu kumpulan *fitur* data. (Maulida, Suyatno, & Hatta, 2016)

2.8 Information Gain (IG)

Information Gain merupakan teknik seleksi *fitur* yang memakai metode *scoring* untuk nominal ataupun pembobotan atribut kontinu yang didiskritkan menggunakan maksimal *entropy*. Suatu *entropy* digunakan untuk mendefinisikan nilai Information Gain. *Entropy* menggambarkan banyaknya informasi yang dibutuhkan untuk mengkodekan suatu kelas. Information Gain (IG) dari suatu *term* diukur dengan menghitung jumlah *bit* informasi yang diambil dari prediksi kategori dengan ada atau tidaknya *term* dalam suatu dokumen. (Maulida, Suyatno, & Hatta, 2016)

Teknik seleksi *fitur* dengan Information gain artinya adalah memilih simpul *fitur* dari pohon keputusan berdasar nilai *information gain*. Nilai *information gain* sebuah *fitur* diukur dari pengaruh *fitur* tersebut terhadap keseragaman kelas pada data yang dipecah menjadi subdata dengan nilai *fitur* tertentu. Keseragaman kelas (*entropy*) dihitung pada data sebelum dipecah dengan persamaan 2.1 dan pada data setelah dipecah dengan persamaan 2.2 berikut ini.

$$Entropy(S) = \sum_{i=1}^k (P_i) \log_2(P_i) \quad (2.1)$$

Dengan nilai P_i adalah proporsi data S dengan kelas i . K adalah jumlah kelas pada *output* S .

$$Entropy(S, A) = \sum_{i=1}^v \left(\frac{S_v}{S} * Entropy(S_v) \right) \quad (2.2)$$

Dengan nilai v adalah semua nilai yang mungkin dari atribut A , S_v adalah subset sari S dimana atribut A bernilai v . Nilai *information gain* dihitung dengan persamaan 2.3 berikut ini:

$$Gain(S, A) = Entropy(S) - Entropy(S, A) \quad (2.3)$$

Dengan nilai $Gain(S, A)$ adalah nilai *information gain*. $Entropy(S)$ adalah nilai *entropy* sebelum pemisah. $Entropy(S, A)$ adalah nilai *entropy* setelah pemisah. Besarnya nilai *information gain* menunjukkan seberapa besar pengaruh suatu atribut terhadap pengklasifikasian data. (Rasywir & Purwarianti, 2015)

2.9 Naïve Bayes Classifier (NB)

Naïve Bayes Classifier adalah salah satu metode yang populer digunakan untuk keperluan *data mining* karena kemudahan penggunaannya (Hall, 2006) serta waktu pemrosesannya yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan tingkat efektifitas yang tinggi (Taheri & Mammadov, 2013).

Dengan bahasa yang lebih sederhana, *Naïve Bayes Classifier* mengasumsikan bahwa keberadaan maupun ketidakberadaan sebuah *fitur* dalam sebuah kelas tidak memiliki keterkaitan dengan keberadaan maupun ketidakberadaan *fitur* lainnya. Sebagai contoh, sesuatu yang berwarna merah, bulat, dan memiliki diameter sekitar 10 cm bisa dikategorikan sebagai buah apel. Walaupun *fitur* ini bergantung antara satu *fitur* dengan *fitur* yang lainnya. *Naïve Bayes Classifier* akan tetap menganggap bahwa *fitur* – *fitur* tersebut independen dan tidak memiliki pengaruh satu sama lainnya (Rocha, 2006).

Bergantung pada model *probabilitas*nya, *Naïve Bayes Classifier* dapat dilatih untuk melakukan *supervised learning* dengan sangat efektif. Dalam berbagai macam penerapannya, estimasi parameter untuk model *Naïve Bayes* menggunakan metode *maximum likelihood*, yang artinya pengguna dapat menggunakan model *Naïve Bayes* tanpa perlu mempercayai *probabilitas Bayesian* atau tanpa menggunakan metode *Bayesian*. (Hadna, Santosa, & Winarno, 2016)

Teorema *Bayes* merupakan teorema yang mengacu pada konsep *probabilitas* bersyarat. Secara umum *teorema Bayes* dapat dinotasikan pada persamaan 2.4 berikut:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)} \quad (2.4)$$

Pada *Naive Bayes Classification* setiap *tweet* direpresentasikan dalam pasangan atribut $(a_1, a_2, a_3, \dots, a_n)$ dimana a_1 adalah kata pertama a_2 adalah kata kedua dan seterusnya, sedangkan V adalah himpunan kelas. Pada saat klasifikasi, metode ini akan menghasilkan kategori / kelas yang paling tinggi *probabilitasnya* (V_{MAP}) dengan memasukkan atribut $(a_1, a_2, a_3, \dots, a_n)$. Adapun rumus V_{MAP} dapat dilihat pada persamaan 2.5 berikut:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2, a_3, \dots, a_n) \quad (2.5)$$

Dengan menggunakan teorema Bayes, maka persamaan (2.5) dapat ditulis menjadi,

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, a_3, \dots, a_n | V_j) P(V_j)}{P(a_1, a_2, a_3, \dots, a_n)} \quad (2.6)$$

$P(a_1, a_2, a_3, \dots, a_n)$ nilainya konstan untuk semua v_j sehingga persamaan (2.6) dapat juga dinyatakan menjadi persamaan 2.7 berikut:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2, a_3, \dots, a_n | V_j) P(V_j) \quad (2.7)$$

Naive Bayes Classifier menyederhanakan hal ini dengan mengasumsikan bahwa didalam setiap kategori, setiap atribut bebas bersyarat satu sama lain. Dengan kata lain,

$$P(a_1, a_2, a_3, \dots, a_n | V_j) = \prod_i P(a_i | v_j) \quad (2.8)$$

Kemudian apabila persamaan (2.7) disubstitusikan ke persamaan (2.8), maka akan menghasilkan persamaan 2.9 berikut:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \times \prod_i P(a_i | v_j) \quad (2.9)$$

$P(v_j)$ dan *probabilitas* kata a_i untuk setiap kategori $P(a_i|v_j)$ dihitung pada saat *training*. Dimana,

$$P(v_j) = \frac{docs_j}{training} \quad (2.10)$$

$$P(a_i|v_j) = \frac{n_i + 1}{n + kosakata} \quad (2.11)$$

Dimana $docs_j$ adalah jumlah dokumen pada kategori j dan *training* adalah jumlah dokumen yang digunakan dalam proses *training*. Sedangkan n_i adalah jumlah kemunculan kata a_i pada kategori v_j , n adalah jumlah kosakata yang muncul pada kategori v_j dan kosakata adalah jumlah kata unik pada semua data *training*. (Rodiysyah & Winarko, 2013)

2.10 Evaluasi

Evaluasi performasi dilakukan untuk menguji hasil dari klasifikasi dengan mengukur nilai performasi dari sistem yang telah dibuat. Parameter pengujian yang digunakan untuk evaluasi yaitu akurasi yang perhitungannya dari tabel *confusion matrix* (matriks klasifikasi atau tabel *kontigensi*). Tabel 2.1 menampilkan sebuah *confusion matrix* untuk pengklasifikasian kedalam dua kelas. (Novantirani, Sabariah, & Effendy, 2015)

Tabel 2.1 Confusion Matrix

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Matriks tersebut memiliki empat nilai yang dijadikan acuan dalam perhitungan, yaitu :

- True Positive* (TP), ketika kelas yang diprediksi positif dan faktanya positif.
- True Negative* (TN), ketika kelas yang diprediksi negatif, dan faktanya negatif.

- c) *False Positive* (FP), ketika kelas yang diprediksi positif dan faktanya negatif.
- d) *False Negative* (FN), ketika kelas yang diprediksi negatif dan faktanya positif.

Alternatif yang jelas terlintas pada pikiran pembaca dalam menilai sebuah sistem adalah dengan akurasi. Akurasi adalah ketepatan suatu sistem melakukan klasifikasi yang benar. Perhitungan untuk akurasi dapat dikalkulasi dengan persamaan 2.12 berikut:

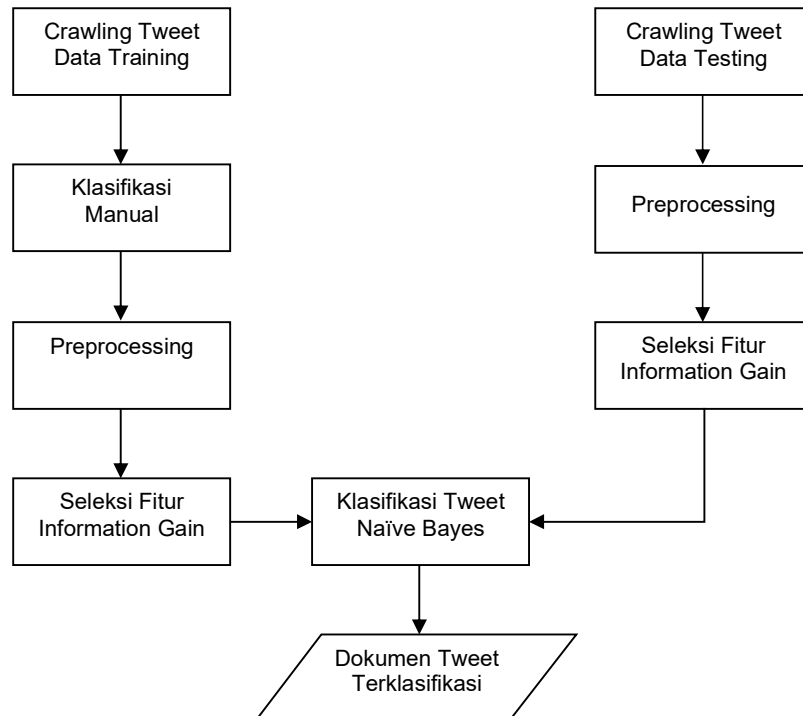
$$Accuracy (A) = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (2.12)$$

BAB III

METODOLOGI PENELITIAN

3.1 Deskripsi Umum

Sistem klasifikasi untuk sentimen analisis yang sedang dikerjakan ini memiliki suatu rancangan bagaimana alur sistem ini akan berjalan. Gambaran umum sistem yang akan dibuat seperti gambar 3.1 sebagai berikut :



Gambar 3.1 Gambaran Umum Sistem


Sistem ini dimulai dengan proses menginput data *tweet* dengan cara *crawling*. Proses *crawling* dibedakan menjadi dua, yaitu *crawling data training* dengan cara manual menggunakan *web browser* dan *fitur inspect element*, dan *crawling data testing* menggunakan *API twitter*. Output dari sistem ini merupakan data *testing* dengan nilai *output* berupa opini positif atau negatif yang diklasifikasikan oleh sistem berdasarkan pembelajaran data *training*.

3.2 Input

Input yang dimasukkan sistem adalah dokumen yang berupa *tweet* dari akun *Twitter* yang berupa opini. Data *tweet* tersebut didapat dengan memanfaatkan *fitur API (Application Interface)* yang telah disediakan oleh *Twitter*. Dokumen

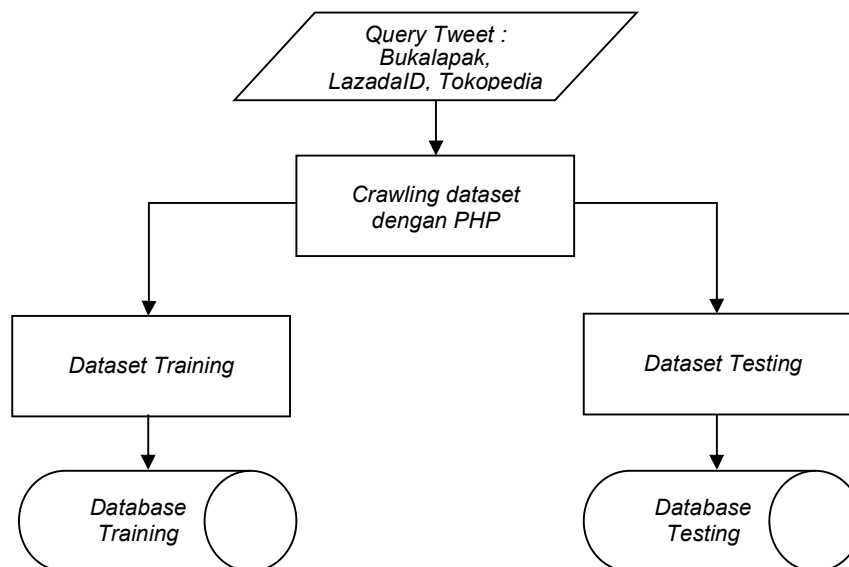
yang dimasukkan merupakan dokumen berbahasa Indonesia, seperti yang ditunjukkan pada table 3.1 berikut :

Tabel 3.1 Dokumen *Tweet*

<i>Tweet</i>	Opini
	P
	N

3.3 Dataset

Dataset berupa teks berbahasa Indonesia yang diambil dari *website* <http://www.twtiiter.com> . Ada beberapa data yang diambil dari *website* tersebut, untuk penelitian ini data yang diambil menggunakan *query* ‘bukalapak’, ‘bukalapak_care’, ‘lazadaid’, ‘lazadaidcare’, ‘tokopedia’, dan ‘tokopediacare’. *Query* tersebut merupakan akun resmi dari *e-commerce* Bukalapak, Lazada, dan Tokopedia. Diagram alur proses pengambilan *dataset* seperti pada gambar 3.2.



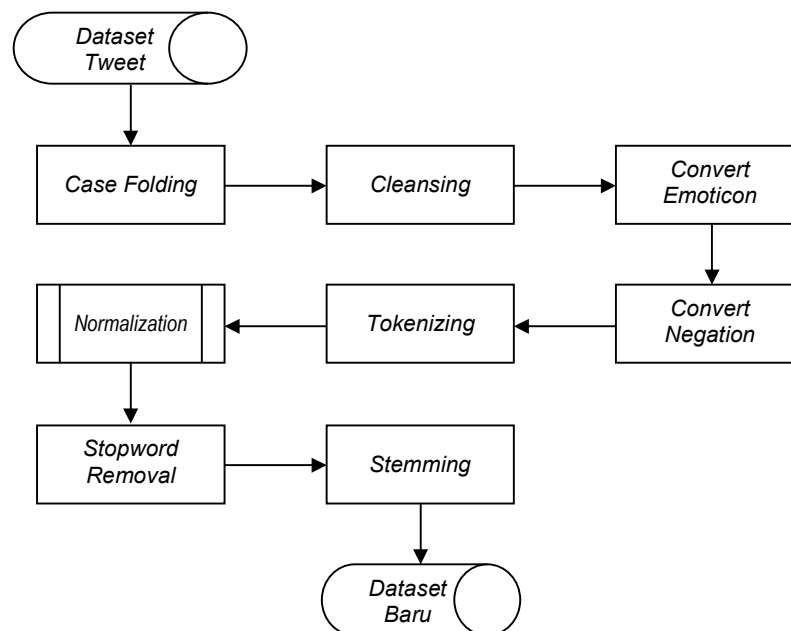
Gambar 3.2 Diagram Proses *Dataset*

Dataset dari hasil *crawling* dengan *API Twitter* ini akan dibagi menjadi dua bagian yaitu data *training* dan data *testing* yang dipresentasikan pada gambar 3.2. Data hasil *crawling* ini berupa dokumen *tweet* yang tidak menyertakan atribut lainnya. Data *training* ini akan dimasukkan ke dalam *database* MySQL, dan diklasifikasikan secara manual dengan label sentimen positif atau negatif. Data *testing* yang diperoleh dari proses *crawling* ini akan disimpan didalam *database* MySQL, yang nantinya akan diolah kedalam sistem untuk menghasilkan *output* otomatis berupa sentimen positif atau negatif.

Pengambilan data *training* dan data *testing* dilakukan dengan cara yang sama, tetapi waktu pengambilannya berbeda. Data *training* diambil selama satu bulan pada bulan maret sedangkan data *testing* diambil selama satu bulan pada bulan april.

3.4 Preprocessing

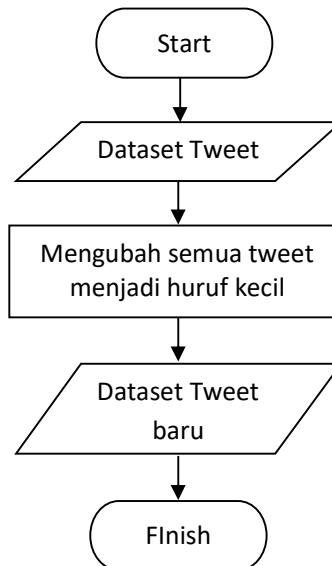
Proses *preprocessing* merupakan hal yang penting untuk tahap selanjutnya, yaitu mengurangi atribut yang tidak berpengaruh terhadap proses klasifikasi. Data yang dimasukkan pada tahap ini masih berupa data mentah yang masih kotor, sehingga hasil dari proses ini adalah dokumen yang berkualitas yang harapannya mempermudah dalam proses klasifikasi. Proses *preprocessing* terdiri dari beberapa tahapan yang dapat dilihat pada gambar 3.3 berikut :



Gambar 3.3 Diagram Alir *Preprocessing*

3.4.1 Case Folding

Pada tahap *case folding* huruf kapital pada semua dokumen *tweet* diubah menjadi huruf kecil semua. Tujuannya untuk menghilangkan *redudansi* data yang hanya berbeda pada hurufnya saja. Berikut diagram alir *case folding* pada Gambar 3.4 .



Gambar 3.4 Diagram Alir Case Folding

Sebagai gambaran dari proses *case folding* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.2.

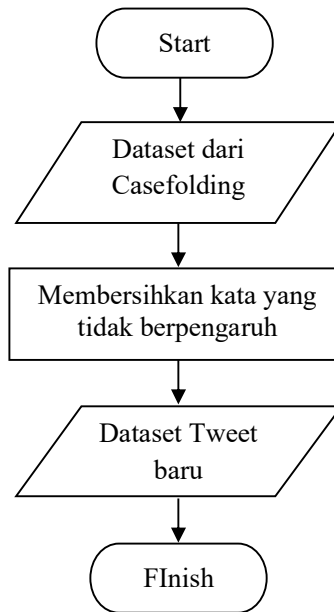
Tabel 3.2 Contoh Tahap Case Folding

<i>Input Process</i>	<i>Output Process</i>
@Yoghaaaaa Hi, pesanan kamu sudah dalam pengiriman. Maksimal kamu menerima produk 16/02/2017. Mohon ditunggu ya. Thanks –Bella	@yoghaaaaa hi, pesanan kamu sudah dalam pengiriman. maksimal kamu menerima produk 16/02/2017. mohon ditunggu ya. thanks –bella

3.4.2 Cleansing

Tahapan *cleansing* merupakan tahap pembersihan kata yang tidak berpengaruh sama sekali terhadap sentimen. Komponen dokumen *tweet* memiliki berbagai atribut yang tidak berpengaruh terhadap sentimen, yaitu *mention* yang diawali dengan atribut ('@'), *hashtag* yang diawali dengan atribut ('#'), *link* yang

diawali dengan atribut ('http','bit.ly') dan karakter simbol (~!@#\$\$%^&*()_+?<>,.?:{}[]). Atribut yang tidak berpengaruh tersebut akan dihilangkan dari dokumen dan akan digantikan dengan karakter spasi. Berikut diagram alir *cleansing* pada Gambar 3.5 .



Gambar 3.5 Diagram Alir *Cleansing*

Sebagai gambaran dari proses *cleansing* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.3.

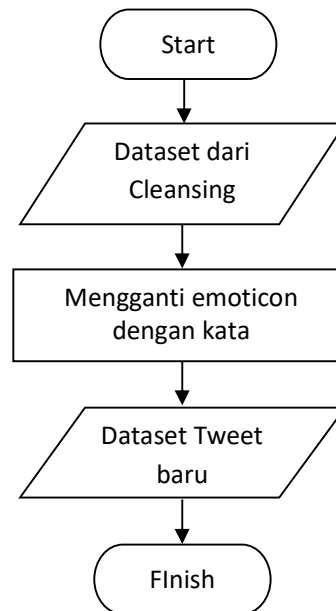
Tabel 3.3 Contoh Tahap *Cleansing*

<i>Input Process</i>	<i>Output Process</i>
@yoghaaaaa hi, pesanan kamu sudah dalam pengiriman. maksimal kamu menerima produk 16/02/2017. mohon ditunggu ya. thanks –bella	hi, pesanan kamu sudah dalam pengiriman maksimal kamu menerima produk 16/02/2017 mohon ditunggu ya thanks –bella

3.4.3 *Convert Emoticon*

Tahap *convert emoticon* ini sangat berpengaruh terhadap sentimen suatu dokumen, karena *emoticon* dapat menggambarkan perasaan seseorang. Karena *emoticon* berupa simbol maka tahap *convert emoticon* ini akan mengubah simbol

senang menjadi kata ‘*emotsenang*’ dan mengubah simbol sedih dengan kata ‘*emotsedih*’. Berikut diagram alir *convert emoticon* pada Gambar 3.6.



Gambar 3.6 Diagram Alir *Convert Emoticon*

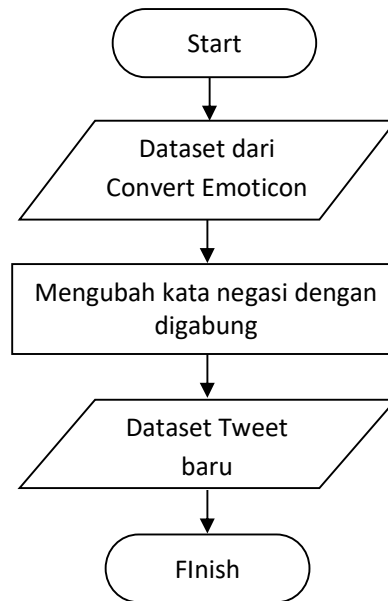
Sebagai gambaran dari proses *convert emoticon* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.4.

Tabel 3.4 Contoh Tahap *Convert Emoticon*

<i>Input Process</i>	<i>Output Process</i>
hi, maaf ya :(bisa infokan atas nama pemesan dan nomor pesanan kamu melalui dm kami ya, agar kami bantu cek –echa	hi, maaf ya emotsedih bisa infokan atas nama pemesan dan nomor pesanan kamu melalui dm kami ya, agar kami bantu cek –echa

3.4.4 *Convert Negation*

Tahap *convert negation* merupakan proses konversi kata – kata negasi yang terdapat pada suatu *tweet*. Kata negasi akan merubah makna sentimen suatu dokumen, sehingga kata negasi akan digabungkan dengan kata selanjutnya. Contoh kata negasi adalah ‘bukan’, ‘tidak’, ‘jangan’. Berikut diagram alir *convert negation* pada Gambar 3.7.



Gambar 3.7 Diagram Alir *Convert Negation*

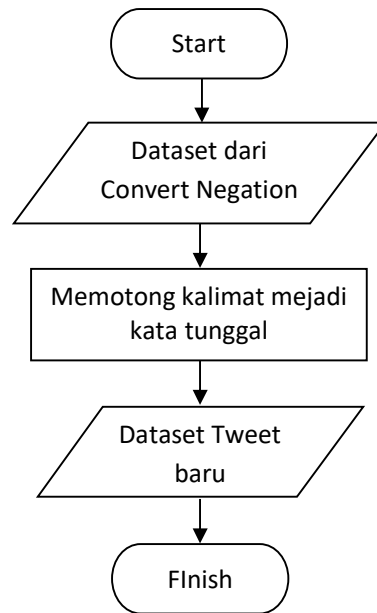
Sebagai gambaran dari proses *convert negation* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.5.

Tabel 3.5 Contoh Tahap *Convert Negation*

<i>Input Process</i>	<i>Output Process</i>
hi, mohon maaf atas ketidaknyamannya, resi yang diberikan pelapak tidak valid, kami sudah melakukan invalid resi (1) ^bl	hi, mohon maaf atas ketidaknyamannya, resi yang diberikan pelapak tidakvalid, kami sudah melakukan invalid resi (1) ^bl

3.4.5 *Tokenizing*

Tahap *tokenizing* merupakan pemotongan kata berdasarkan tiap kata yang menyusunnya menjadi potongan tunggal. Kata dalam dokumen yang dimaksud adalah kata yang dipisah oleh *spasi*. Sehingga hasil dari proses ini merupakan kata tunggal yang dimasukkan ke dalam *database* untuk keperluan pembobotan. Berikut diagram alir *tokenizing* pada Gambar 3.8.



Gambar 3.8 Diagram Alir *Tokenizing*

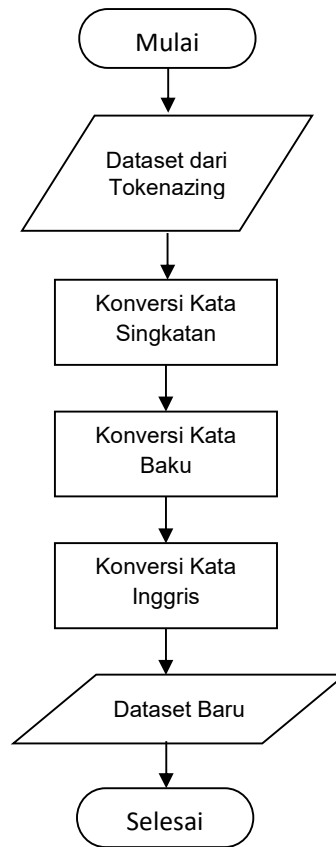
Sebagai gambaran dari proses *tokenizing* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.6.

Tabel 3.6 Contoh Tahap *Tokenizing*

<i>Input Process</i>	<i>Output Process</i>
pesanan sudah dalam pengiriman, jadi sudah tidak bisa dibatalkan ya aga	'pesanan' 'sudah' 'dalam' 'pengiriman' 'jadi' 'sudah' 'tidak' 'bisa' 'dibatalkan' 'ya' 'aga'

3.4.6 *Normalization*

Pada tahap *normalization* ini dilakukan pengubahan kata yang tidak sesuai dengan EYD, sehingga dapat mengurangi hasil sentimen dokumen. Tahap ini dibagi menjadi tiga langkah, yaitu konversi kata singkatan, konversi kata baku, dan konversi kata inggris. Berikut diagram alir *normalization* pada Gambar 3.9.



Gambar 3.9 Diagram Alir *Normalization*

Pengguna *twitter* dibatasi hanya bisa melakukan posting 140 karakter saja, sehingga menyebabkan banyak pengguna yang menulis dengan kata singkatan agar apa yang ditulisnya dapat terekspresikan. Hal tersebut mebuat masalah terhadap performasi sentimen dokumen. Pada tahap ini dilakukan proses untuk mengubah kata singkatan menjadi kata normal, yang dicontohkan pada tabel 3.7.

Tabel 3.7 Contoh Tahap Konversi Kata Singkatan

<i>Input Process</i>	<i>Output Process</i>
hi, bukalapak utk sementara tdk dpt diakses & masih dalam proses utk memulihkannya mohon maaf atas ketidaknyamanannya ^bl	hi bukalapak untuk sementara tidak dapat diakses amp masih dalam proses untuk memulihkannya mohon maaf atas ketidaknyamanannya bl

Proses selanjutnya adalah tahap konversi kata baku, yang berguna untuk mengubah kata yang tidak standar bahasa Indonesia menjadi bahasa Indonesia yang baku. Biasanya pengguna *twitter* masih menggunakan bahasa daerahnya dan

bahasa gaul sehingga perlu dilakukan konversi kata baku, seperti yang dicontohkan pada tabel 3.8.

Tabel 3.8 Contoh Tahap Konversi Kata Baku

<i>Input Process</i>	<i>Output Process</i>
hai yunni mohon maaf mohon informasikan terlebih dahulu email akun tokopedia via dm	hai yunni mohon maaf mohon informasi lebih dahulu email akun via

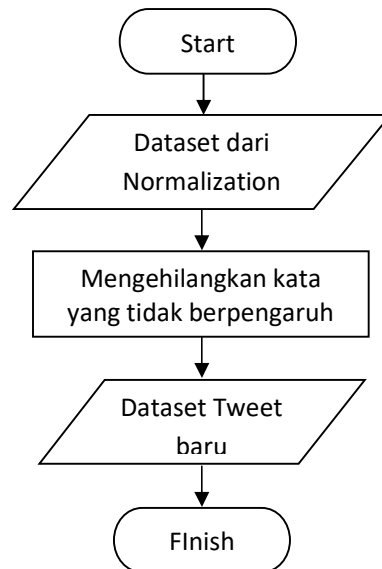
Tahap yang terakhir dalam proses *normalization* adalah mengubah kata Inggris ke Indonesia, berikut contoh konversi kata Inggris pada tabel 3.9.

Tabel 3.9 Contoh Tahap Konversi Kata Inggris

<i>Input Process</i>	<i>Output Process</i>
ulasan customer service yang menggunakan gosend hari ini thanks buat bintang nya	ulasan customer service yang menggunakan gosend hari ini terimakasih buat bintang nya

3.4.7 Stopword Removal

Tahap *stopword removal* merupakan tahap menghilangkan kata yang tidak sesuai dengan topik dokumen, yang jika ada kata tersebut tidak mempengaruhi akurasi dalam klasifikasi sentimen dokumen. Kata yang akan dihilangkan dihipunkan dalam *database* kata *stopword*. Jika dalam dokumen *tweet* ada yang sesuai dengan kata dalam *stopword* maka kata tersebut akan dihilangkan dan diganti dengan karakter spasi. Berikut diagram alir *stopword removal* pada Gambar 3.10.



Gambar 3.10 Diagram Alir *Stopword Removal*

Sebagai gambaran dari proses *stopword removal* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.10.

Tabel 3.10 Contoh Tahap *Stopword Removal*

<i>Input Process</i>	<i>Output Process</i>
bukan kah resinya sudah saya kirim dan juga sudah di konfirm oleh lazada klo barangnya sudah sampai di gudang lazada	bukan resinya kirim konfirm lazada klo barangnya gudang lazada

3.4.8 *Stemming*

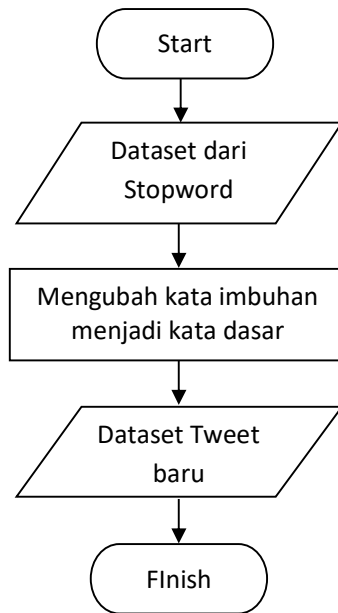
Pada tahap *stemming* merupakan suatu proses untuk mengubah kata – kata yang terdapat dalam suatu dokumen ke kata – kata akarnya dengan menggunakan aturan – aturan tertentu. Proses *stemming* bahasa Indonesia dengan menghilangkan *sufiks*, *prefix*, dan *konfiks* pada dokumen. Pada proses *stemming* ini penulis menggunakan metode yang dibuat oleh Bobby Nazief dan Mirna Adriani, dengan tahapan sebagai berikut: (Agusta, 2009)

- 1) Cari kata yang akan distem dalam kamus. Jika ditemukan maka diasumsikan bahwa kata tersebut adalah *root word*. Maka metode berhenti.
- 2) *Inflection Suffixes* (“-lah”, “-kah”, “-ku”, “-mu”, atau “-nya”) dibuang. Jika berupa *particles* (“-lah”, “-kah”, “-tah” atau “-pun”) maka langkah ini

diulangi lagi untuk menghapus *Possesive Pronouns* (“-ku”, “-mu”, atau “-nya”), jika ada.

- 3) Hapus *Derivation Suffixes* (“-i”, “-an” atau “-kan”). Jika kata ditemukan di kamus, maka metode berhenti. Jika tidak maka ke langkah 3a
 - a. Jika “-an” telah dihapus dan huruf terakhir dari kata tersebut adalah “-k”, maka “-k” juga ikut dihapus. Jika kata tersebut ditemukan dalam kamus maka metode berhenti. Jika tidak ditemukan maka lakukan langkah 3b.
 - b. Akhiran yang dihapus (“-i”, “-an” atau “-kan”) dikembalikan, lanjut ke langkah 4.
- 4) Hapus *Derivation Prefix*. Jika pada langkah 3 ada sufiks yang dihapus maka pergi ke langkah 4a, jika tidak pergi ke langkah 4b.
 - a. Periksa tabel kombinasi awalan – akhiran yang tidak diijinkan. Jika ditemukan maka metode berhenti, jika tidak pergi ke langkah 4b.
 - b. *For i = 1 to 3*, tentukan tipe awalan kemudian hapus awalan. Jika *root word* belum juga ditemukan lakukan langkah 5, jika sudah maka metode berhenti. Catatan: jika awalan kedua sama dengan awalan pertama metode berhenti.
- 5) Melakukan *recording*.
- 6) Jika semua langkah telah selesai tetapi tidak juga berhasil maka kata awal diasumsikan sebagai *root word*. Proses selesai.

Berikut diagram alir *stemming* pada Gambar 3.11.



Gambar 3.11 Diagram Alir *Stemming*

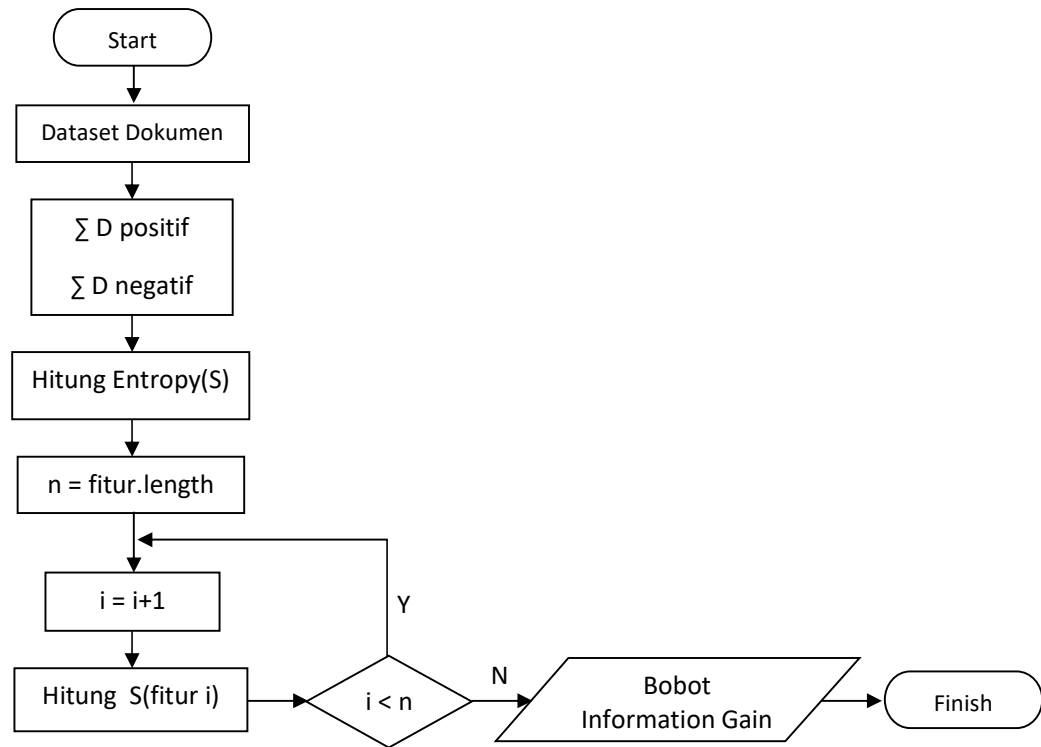
Sebagai gambaran dari proses *stemming* berikut contoh *tweet* yang dihasilkan seperti pada tabel 3.11.

Tabel 3.11 Contoh Tahap *Stemming*

<i>Input Process</i>	<i>Output Process</i>
kalau tetap tidak terkirim pakatnya dianggap batal gimana kalau batal pesanan min pembayaran via cod	kalau tetap tidak kirim paket anggap batal gimana pesan kalau batal pesan min bayar via cod

3.4.9 Seleksi Fitur *Information Gain*

Pada proses Seleksi Fitur ini merupakan proses untuk penyeleksian *fitur* yang paling penting dalam proses klasifikasi. Berikut *flowchart Information Gain* pada gambar 3.12.



Gambar 3.12 Flowchart Inforamtion Gain

Dengan cara menghitung bobot sesuai rumus 2.3 Pada bab 2, metode *information gain* secara sederhana dapat dicontohkan seperti tabel 3.12.

Tabel 3.12 Contoh Koleksi Data

Dokumen	Fitur			Sentimen
	Puas	Terimakasih	Batal	
D ₁	Ya	Ya	Tidak	P
D ₂	Ya	Ya	Tidak	P
D ₃	Ya	Ya	Tidak	P
D ₄	Tidak	Ya	Ya	N
D ₅	Tidak	Tidak	Ya	N
D ₆	Tidak	Tidak	Ya	N
D ₇	Tidak	Tidak	Tidak	P
D ₈	Ya	Tidak	Tidak	N
D ₉	Ya	Ya	Ya	P
D ₁₀	Ya	Ya	Tidak	P

Fitur yang terdapat pada tabel 3.12 merupakan potongan kata dari dokumen yang akan kita hitung bobotnya. Contoh kasus perhitungan bobot *information gain* pada kata ‘puas’ dengan menghitung *entropy* pada *dataset* dengan menggunakan persamaan 2.1, sebagai berikut:

$$Entropy (Set) = - \left[\left(\frac{6}{10} \right) \log_2 \left(\frac{6}{10} \right) + \left(\frac{4}{10} \right) \log_2 \left(\frac{4}{10} \right) \right] = 0,971$$

Selanjutnya ambil contoh pada kata ‘puas’ yang memiliki *value* Ya atau Tidak sehingga dapat dihitng dengan persamaan 2.1 setelah itu hitung *Entropy* (S_{puas}) dengan persamaan 2.2 dan menghasilkan perhitungan sebagai berikut :

$$Entropy (Ya) = - \left[\left(\frac{4}{6} \right) \log_2 \left(\frac{4}{6} \right) + \left(\frac{2}{6} \right) \log_2 \left(\frac{2}{6} \right) \right] = 0,9183$$

$$Entropy (Tidak) = - \left[\left(\frac{2}{4} \right) \log_2 \left(\frac{2}{4} \right) + \left(\frac{2}{4} \right) \log_2 \left(\frac{2}{4} \right) \right] = 1$$

$$Entropy (S_{puas}) = \left(\frac{6}{10} \right) 0,9183 + \left(\frac{4}{10} \right) \times 1 = 0,95098$$

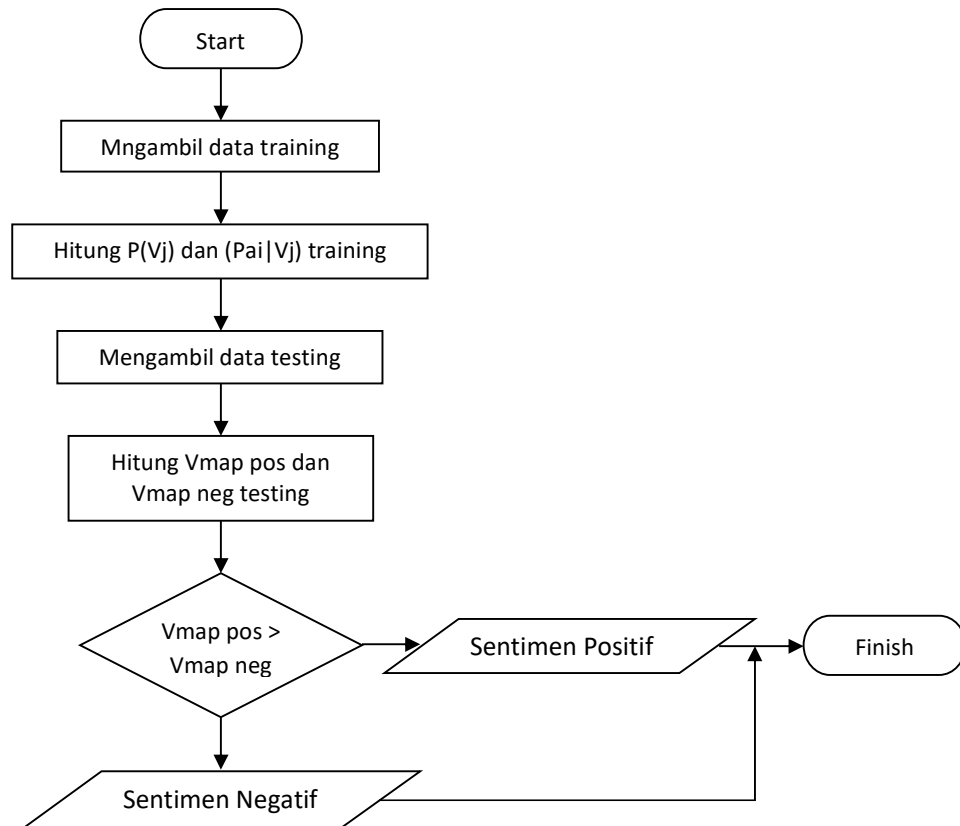
Langkah terakhir untuk mencari nilai *information gain* menggunakan persamaan 2.3 sebagai berikut :

$$Gain(S_{puas}) = 0,971 - 0,95098 = 0,02002$$

Dengan kata lain *information gain* merupakan bentuk nilai yang menunjukkan seberapa besar pengaruh suatu *fitur* terhadap pengklasifikasian data. Sehingga hasil dari *information gain* yang tinggi yang akan dilolah kedalam sistem klasifikasi.

3.5 Naïve Bayes Classification

Pada tahap klasifikasi ini menggunakan metode *Naïve Bayes* dibagi menjadi dua proses, yaitu proses *training* dan proses *testing*. Pada tahap ini dilakukan proses *training* terlebih dahulu untuk pelatihan, selanjutnya dilakukan proses *testing* dengan mengacu *probabilitas* dari *dataset training*. Alur tahapan klasifikasi *Naïve Bayes* ini ditunjukkan pada gambar 3.13 berikut.



Gambar 3.12 Flowchart Naïve Bayes

Berikut penulis membuat contoh perhitungan *Naïve Bayes Classification* dengan sampel 6 *tweet* :

1. Proses *Training*

Sebuah dokumen *training* sudah diklasifikasikan secara manual dan sudah dilakukan proses *preprocessing* seperti pada tabel 3.13 berikut:

Tabel 3.13 Contoh Kasus Data Training

Tweet	Fitur	Kategori
Tweet1	sudah kami bantu remit silakan cek lebih lanjut	P
Tweet2	baik, tunggu. terimakasih banyak	P
Tweet3	silakan kami sudah respon pesan terimakasih	P
Tweet4	mohon maaf, resi tidak valid	N
Tweet5	tidak kenal, email tidak ada	N
Tweet6	mohon maaf, mohon informasi	N

Dari data tabel 3.13 dibuat sebuah model *probabilitas* dengan mengacu pada persamaan 2.10 dan 2.11 sebagai berikut:

$$P(a_{\text{terimakasih}} | V_{\text{positif}}) = \frac{2 + 1}{18 + 32} = \frac{3}{50}$$

$$P(a_{\text{terimakasih}} | V_{\text{negatif}}) = \frac{0 + 1}{14 + 32} = \frac{1}{46}$$

Jika dibuat menjadi sebuah tabel, maka *probabilitas* setiap kata pada data *training* seperti pada tabel 3.14 berikut:

Tabel 3.14 Perhitungan Probabilitas Data Training

Kategori	P(v _j)	P(a _i V _j)							
		sudah	kami	bantu	remit	silakan	cek	lebih	lanjut
P	$\frac{1}{2}$	$\frac{3}{50}$	$\frac{3}{50}$	$\frac{2}{50}$	$\frac{2}{50}$	$\frac{3}{50}$	$\frac{2}{50}$	$\frac{2}{50}$	$\frac{2}{50}$
N	$\frac{1}{2}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$

Kategori	P(v _j)	P(a _i V _j)						
		Baik	tunggu	terimakasih	banyak	respon	pesan	mohon
P	$\frac{1}{2}$	$\frac{2}{50}$	$\frac{2}{50}$	$\frac{3}{50}$	$\frac{2}{50}$	$\frac{2}{50}$	$\frac{2}{50}$	$\frac{1}{50}$
N	$\frac{1}{2}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{1}{46}$	$\frac{4}{46}$

Kategori	P(v _j)	P(a _i V _j)							
		maaf	resi	tidak	valid	kenal	Email	ada	informasi
P	$\frac{1}{2}$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{1}{50}$	$\frac{1}{50}$
N	$\frac{1}{2}$	$\frac{3}{46}$	$\frac{2}{46}$	$\frac{4}{46}$	$\frac{2}{46}$	$\frac{2}{46}$	$\frac{2}{46}$	$\frac{2}{46}$	$\frac{2}{46}$

Hasil perhitungan *probabilitas* tersebut digunakan sebagai model probabilistik yang selanjutnya digunakan sebagai data acuan untuk menentukan data *testing*.

2. Proses *Testing*

Berikut penulis beri contoh perhitungan untuk data tester, dengan data tester seperti pada tabel 3.15.

Tabel 3.15 Contoh Kasus Data Testing

Tweet	Fitur	Kategori
Tweet7	terimakasih telah pesan barang lapak	?
Tweet8	silakan tunggu hadiah telah pesan	?
Tweet9	maaf tidak bisa kirim besok	?

Proses *testing* ini dihitung *probabilitasnya* dan dicari *probabilitas* tertinggi menggunakan persamaan 2.9 sebagai berikut:

$$\begin{aligned}
 &P(\text{Tweet7}|V_{positif}) \\
 &= P(a_{\text{terimakasih}}|V_{positif}) \times P(a_{\text{telah}}|V_{positif}) \times P(a_{\text{pesan}}|V_{positif}) \times \\
 &\quad P(a_{\text{barang}}|V_{positif}) \times P(a_{\text{lapak}}|V_{positif}) \times P(V_{positif}) \\
 &= \frac{3}{50} \times 1 \times \frac{2}{50} \times 1 \times 1 \times \frac{1}{2} \\
 &= 0,0012
 \end{aligned}$$

$$\begin{aligned}
 &P(\text{Tweet7}|V_{negatif}) \\
 &= P(a_{\text{terimakasih}}|V_{negatif}) \times P(a_{\text{telah}}|V_{negatif}) \times P(a_{\text{pesan}}|V_{negatif}) \times \\
 &\quad P(a_{\text{barang}}|V_{negatif}) \times P(a_{\text{lapak}}|V_{negatif}) \times P(V_{negatif}) \\
 &= \frac{1}{46} \times 1 \times \frac{1}{46} \times 1 \times 1 \times \frac{1}{2} \\
 &= 0,000236295
 \end{aligned}$$

$$\begin{aligned}
 &P(\text{Tweet8}|V_{positif}) \\
 &= P(a_{\text{silakan}}|V_{positif}) \times P(a_{\text{tunggu}}|V_{positif}) \times P(a_{\text{hadiah}}|V_{positif}) \times \\
 &\quad P(a_{\text{telah}}|V_{positif}) \times P(a_{\text{pesan}}|V_{positif}) \times P(V_{positif}) \\
 &= \frac{3}{50} \times \frac{2}{50} \times 1 \times 1 \times \frac{2}{50} \times \frac{1}{2} \\
 &= 0,000048
 \end{aligned}$$

$$\begin{aligned}
& P(Tweet8|V_{negatif}) \\
&= P(a_{silakan}|V_{negatif}) \times P(a_{tunggu}|V_{negatif}) \times P(a_{hadia}|V_{negatif}) \times \\
&\quad P(a_{telah}|V_{negatif}) \times P(a_{pesan}|V_{negatif}) \times P(V_{negatif}) \\
&= \frac{1}{46} \times \frac{1}{46} \times 1 \times 1 \times \frac{1}{46} \times \frac{1}{2} \\
&= 0,000005137
\end{aligned}$$

$$\begin{aligned}
& P(Tweet9|V_{positif}) \\
&= P(a_{maaf}|V_{positif}) \times P(a_{tidak}|V_{positif}) \times P(a_{bisa}|V_{positif}) \times \\
&\quad P(a_{kirim}|V_{positif}) \times P(a_{besok}|V_{positif}) \times P(V_{positif}) \\
&= \frac{1}{50} \times \frac{1}{50} \times 1 \times 1 \times 1 \times \frac{1}{2} \\
&= 0,00002
\end{aligned}$$

$$\begin{aligned}
& P(Tweet9|V_{negatif}) \\
&= P(a_{maaf}|V_{negatif}) \times P(a_{tidak}|V_{negatif}) \times P(a_{bisa}|V_{negatif}) \times \\
&\quad P(a_{kirim}|V_{negatif}) \times P(a_{besok}|V_{negatif}) \times P(V_{negatif}) \\
&= \frac{3}{46} \times \frac{2}{46} \times 1 \times 1 \times 1 \times \frac{1}{2} \\
&= 0,001417769
\end{aligned}$$

Setelah menghitung *probabilitas* dari setiap data tester, diperoleh hasil pada tabel 3.16 sebagai berikut:

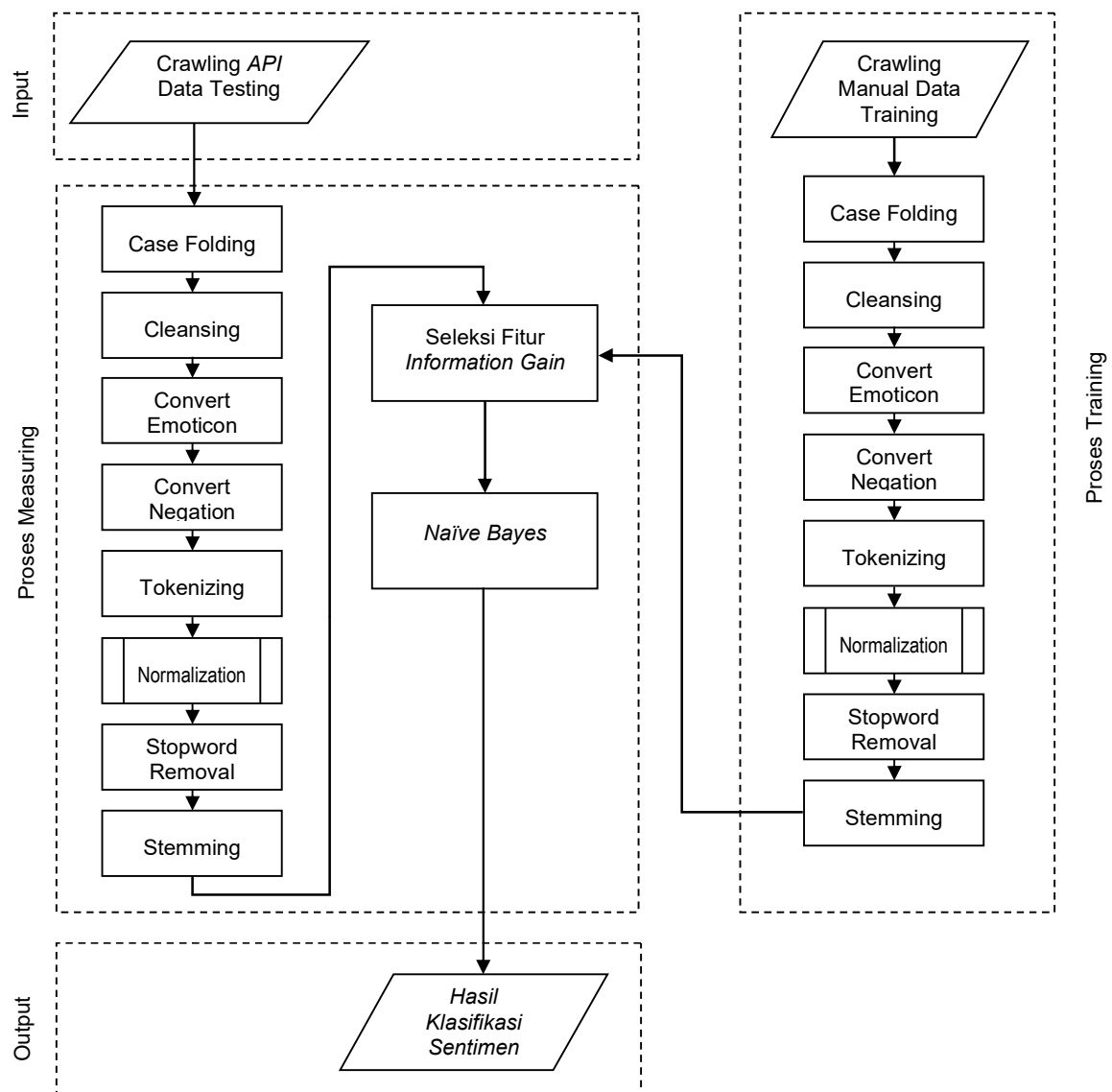
Tabel 3.16 Nilai Probabilitas Data Testing

Tweet	Probabilitas	
	Positif	Negatif
Tweet7	0,0012	0,000236295
Tweet8	0,000048	0,000005137
Tweet9	0,00002	0,001417769

Pada tabel 3.16 kita dapat menganalisis hasil dari data *testing* pertama yaitu Tweet7 didapatkan nilai *probabilitas* positif lebih besar dari nilai *probabilitas*

negatif sehingga dapat disimpulkan bahwa *Tweet7* termasuk kategori sentimen positif. Pada *Tweet8* didapatkan nilai *probabilitas* positif lebih besar dari nilai *probabilitas* negatif sehingga *Tweet8* termasuk kategori sentimen positif. Sedangkan pada *Tweet9* nilai *probabilitas* negatif lebih besar dari nilai *probabilitas* positif sehingga *Tweet9* termasuk kategori sentimen negatif.

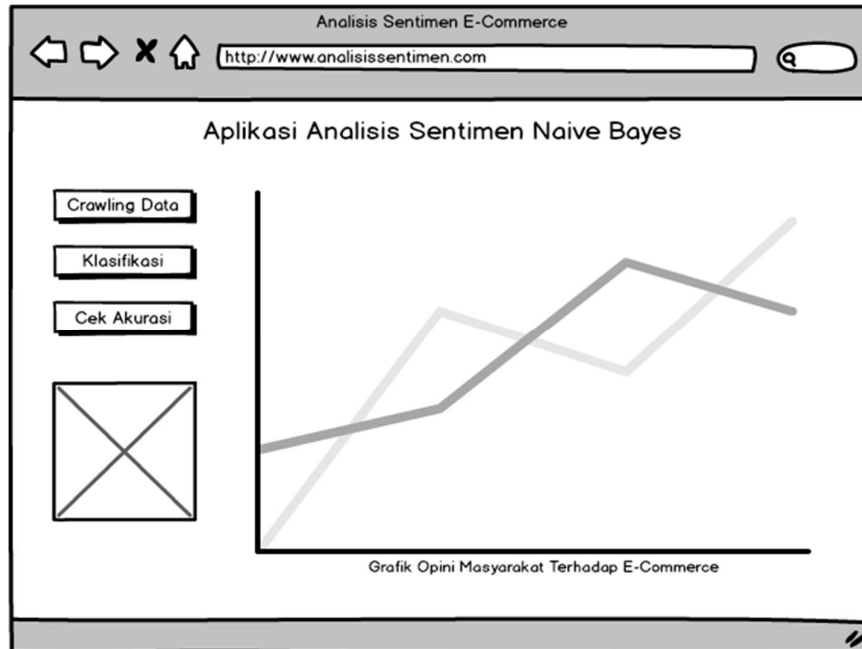
3.6 Perancangan Desain Sistem



Gambar 3.12 Diagram Alir Sistem

3.7 Hasil Analisis Sistem

Sistem ini akan menghasilkan suatu grafik opini yang menampilkan data *testing*. Tampilan ini berupa tampilan GUI yang isinya meliputi hasil sentimen positif dan negatif, serta juga akurasi dari sistem klasifikasi metode *Naïve Bayes*. Rancangan *mockup GUI* yang penulis buat seperti gambar 3.5 berikut:



Gambar 3.13 Rancangan Tampilan GUI

DAFTAR PUSTAKA

- Aditya, B. R. (2015). Penggunaan Web Crawler untuk Menghimpun Tweet dengan Metode Pre-Processing *Text Mining*. *Jurnal Infotel Vol. 7 No. 2* .
- Aditya, C. S., Hani'ah, M., Fitrawan, A. A., Arifin, A. Z., & Purwitasari, D. (2016). Deteksi Bot Spammer pada *Twitter* Berbasis Sentiment Analysis. 7.
- Agusta, L. (2009). Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem dan Informatika (KNSN)* .
- Chandani, V., Wahono, R. S., & Purwanto. (2015). Komparasi Algoritma Klasifikasi Machine Learning Dan Feature. *Journal of Intelligent Systems* .
- Faradhillah, N. Y., Kusumawardani, R. P., & Hafidz, I. (2016). Eksperimen Sistem Klasifikasi Analisa Sentimen *Twitter* pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin. *Seminar Nasional Sistem Informasi Indonesia* .
- Hadna, N. M., Santosa, P. I., & Winarno, W. W. (2016). Studi Literatur tentang Perbandingan Metode untuk Proses Analisis Sentimen di *Twitter*. *Seminar Teknologi Informasi dan komunikasi (SENTIKA)* .
- Hermawati, F. A. (2013). *Data Mining*. Yogyakarta: Penerbit Andi.
- Kurniawan, B., Effendi, S., & Sitompul, O. S. (2012). Klasifikasi Konten Berita Dengan Metode *Text Mining*. *JURNAL DUNIA TEKNOLOGI INFORMASI Vol. 1, No. 1* , 14 - 19.
- Maulida, I., Suyatno, A., & Hatta, H. R. (2016). Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain. *JSM STMIK* .
- Mujilahwati, S. (2016). Pre-processing *Text Mining* pada *Twitter*. *Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA)* .
- Novantirani, A., Sabariah, M. K., & Effendy, V. (2015). Analisis Sentimen pada *Twitter* untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine. *Program Studi Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung* .

- Nurhuda, F., Sihwi, S. W., & Doewes, A. (2013). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari *Twitter* menggunakan Metode Naive Bayes Classifier. *Jurnal ITSMART*.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 6.
- Pang, B., & Lee, L. (2002). Thumbs up? Sentiment Classification using Machine Learning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia: Association for Computational Linguistics.
- Putranti, D. N., & Winarko, E. (2014). Analisis Sentimen *Twitter* untuk Teks Berbahasa. *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*.
- Rasywir, E., & Purwarianti, A. (2015). Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. *Jurnal Cybermatika*.
- Rodiyansyah, S. F., & Winarko, E. (2013). Klasifikasi Posting *Twitter* Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*.
- Rozi, I. F., Pranomo, S. H., & Dahlan, E. A. (2012). Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. *Electrics, Electronics, Communications, Controls, Informatics, Systems (EECCIS)*.
- Saadah, M. N., Atmagi, R. W., Rahayu, D. S., & Arifin, A. Z. (2013). Sistem Temu Kembali Dokumen Teks dengan Pembobotan TF-IDF dan LCS. *Jurnal Ilmiah Teknologi Informasi (JUTI)*.
- Wati, R. (2016). Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan menggunakan Naive Bayes. *Jurnal Evolusi*.