Analisis Sentimen Layanan Provider Telepon Seluler pada Twitter menggunakan Metode Naïve Bayesian Classification

Ronny Julianto, Evi Dianti Bintari, Indrianti

Abstraksi— **Pengguna twitter saat ini banyak menulis tentang opini-opini yang mengandung sifat positif, negatif, maupun netral terhadap suatu produk ataupun layanan jasa dengan menggunakan singkatan kata dan ejaan kata yang singkat serta tidak tepat sehingga menimbulkan salah penafsiran pendapat yang akan disampaikan. Oleh sebab itu, penelitian ini mencoba menganalisis tweet pada layanan provider telepon seluler sehingga dapat mempercepat proses klasifikasi dan mendapatkan kategori sentimen yang sesuai.**

Analisis dilakukan dengan mengklasifikasikan tweet yang berisi sentimen masyarakat terhadap layanan provider telepon seluler tertentu. Analisis ini melewati tahap *text preprocessing* yang terdiri dari *case folding*, normalisasi fitur, *convert emoticon*, *tokenizing*, normalisasi kata, *stopword removal* dan *stemming*. Metode klasifikasi yang digunakan dalam penelitian ini adalah *Naive Bayesian Classification*. *Naive Bayes Classification* digabungkan dengan fitur untuk dapat *convert emoticon*, serta mengubah kata gaul menjadi kata baku dan menghitung akurasi menggunakan *confusion matrix*.

Tahapan klasifikasi tweet berdasarkan data pelatihan yang telah diketahui kategorinya dan proses klasifikasi dokumen yang belum diketahui kategorinya. Penulis menggunakan 600 data tweet berkaitan dengan sentimen provider telepon seluler. Dari uji coba analisis sentimen serta klasifikasi yang dilakukan terhadap beberapa data tweet provider telepon seluler, akurasi yang dihasilkan mencapai 74%.

***Kata Kunci*—Analisis Sentimen, Twitter, Provider, Naïve Bayesia Classification**

I. PENDAHULUAN

Perkembangnya teknologi informasi yang begitu pesat ini, banyak memberikan dampak positif maupun negatif khususnya di Indonesia. Dampak positifnya seperti, masyarakat Indonesia sekarang ini sangat mudah dalam mengakses informasi-informasi di seluruh dunia tanpa kendala jarak dengan menggunakan teknologi yang disebut dengan *internet*. Hal ini, secara tidak langsung, membuat masyarakat Indonesia menjadi pintar karena mempunyai wawasan yang luas. Tetapi dampak negatifnya, masyarakat perlu menyaring informasi-informasi yang masuk, karena belum tentu isi dari sebuah informasi tersebut benar dan berguna. Belum lagi, saat ini sudah ada wadah yang bisa digunakan untuk menjadi perpanjangan "lidah manusia" dalam menyampaikan suatu pendapat atau pandangan yang disebut dengan media sosial.

Beberapa contoh media sosial yang sedang populer di Indonesia yaitu *Facebook, Twitter, Youtube, Google+*, dsb. Khususnya media sosial yang bernama *Twitter*, sekarang ini sedang digandrungi oleh generasi muda Indonesia karena layanan *microblogging* tersebut *user friendly* sehingga pengguna mudah menggunakannya, terutama dalam menyampaikan pesan atau tweet. Kemudian yang menjadi kelebihan Twitter, yaitu tidak membatasi pertemanan yang biasa disebut *followers*. Penulis pesan tersebut menulis tentang kehidupan penulis, berbagi opini tentang berbagai topik dan membahas isu-isu yang terjadi pada saat ini.

Tingginya pengguna Twitter sendiri menjadi bukti bahwa media sosial menjadi jalur utama untuk persaingan antara tiap operator untuk memenuhi kebutuhan internet pun terus menerus meningkat dengan adanya media sosial seperti Twitter. Jumlah pengguna Twitter lewat media telepon genggam pun lebih banyak daripada lewat komputer maupun laptop, yang membutuhkan fasilitas internet maupun sms yang diberikan oleh operator telepon untuk mengakses media sosial tersebut.

Perusahaan yang bergerak di bidang layanan operator telepon untuk mengetahui kualitas layanan yang telah diberikan kepada konsumen lewat *text mining* menggunakan penilaian opini publik dari media sosial Twitter. Dalam proses pengerjaannya, penulis akan menganalisis data mengenai opini konsumen kepada provider telepon dan menilai kualitas pelayanan dari perusahaan. Hasil penilaian ini mendeskripsikan perusahaan manakah yang telah memberikan servis terbaik dibandingkan perusahaan lainnya yang dapat digunakan sebagai sebuah informasi bagi konsumen untuk menentukan pilihan dalam menggunakan layanan dari salah satu perusahaan provider telepon.

Akun Twitter yang dipilih untuk mempresentasikan produk perusahaan provider telepon terdiri dari, akun Telkomsel diambil dari halaman www.telkomsel.com, www.twitter.com/Kartu_As dan www.twitter.com/simPATI. Akun Indosat diambil dari www.indosat.com/Personal yang menunjukkan ketiga akun lewat www.twitter.com/IndosatMania. Akun XL Axiata diambil dari www.xl.co.id yang menunjukkan www.twitter.com/XL123, www.twitter.com/XLCare dan www.twitter.com/XLandMe.

Sentiment Analysis merupakan salah satu studi komputasional dari opini, appraisal dan emosi melalui entitas, event dan atribut yang dimiliki. Analisis sentimen pada Twitter terdapat kelemahan dalam kata-kata yang terdapat pada kalimat

yang diposting oleh pengguna situs tersebut. Twitter hanya memungkinkan pengguna menulis sebanyak 140 karakter, hal ini yang menyebabkan para pengguna sering menggunakan singkatan kata dan ejaan kata yang salah.

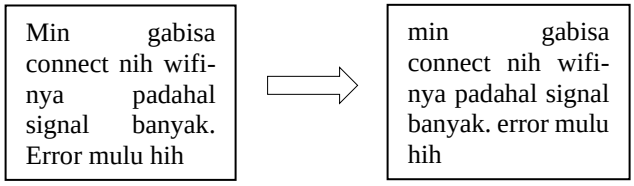
Cara penulisan yang salah tersebut mengakibatkan terjadi kelemahan para proses text mining, dimana dapat menyulitkan fitur yang diambil serta mengurangi ketepatan klasifikasi. Oleh karena itu disini penulis akan menggunakan metode *uni-gram* karakter kata untuk mengambil fitur-fitur yang ada pada sebuah tweet yang kemudian akan diklasifikasi dengan metode *Naive Bayes Classifier*.

II. TINJAUAN PUSTAKA

Text Mining (penambangan teks) adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen word, pdf, kutipan teks dan lain-lain [1].

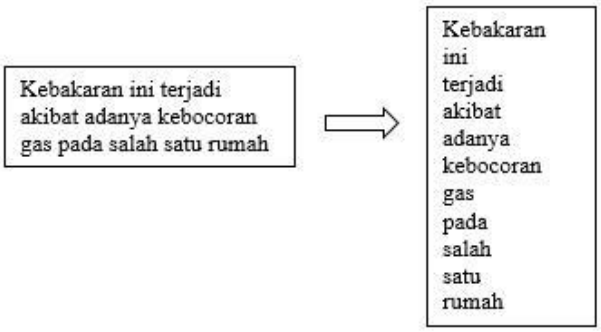
Tahap *text preprocessing* merupakan tahap awal dari text mining. Text preprocessing merupakan proses menggali, mengolah, mengatur informasi dengan cara menganalisis hubungannya, aturan-aturan yang ada di data tekstual semi terstruktur atau tidak terstruktur [2]. Untuk lebih efektif dalam proses dilakukan langkah transformasi data ke dalam suatu format yang memudahkan untuk kebutuhan pengguna, proses ini disebut *preprocessing* dokumen. Setelah dalam bentuk yang lebih terstruktur dengan adanya proses data tersebut sehingga dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut. Tahapan dalam text preprocessing terdiri dari *case folding*, *tokenizing*, normalisasi fitur, *stopword removal* dan *stemming*. Pada proses preprocessing analisis sentimen,ada tambahan tahapan seperti *convert emoticon* dan normalisasi kata.

Case folding adalah tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil, hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf tersebut dihilangkan dan dianggap sebagai pembatas. Contoh dari case folding dapat dilihat pada gambar 2.1.



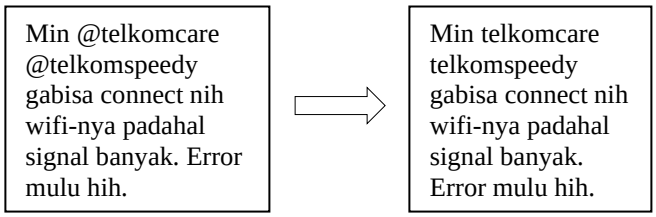
Gambar 2.1 Contoh Case Folding

Tokenizing merupakan sebuah proses yang dilakukan untuk menjadikan sebuah kalimat menjadi lebih bermakna dengan cara memecah kalimat tersebut menjadi kata. Proses ini melakukan penguraian deskripsi yang semula berupa kalimat berisi kata-kata dan tanda pemisah antara kata seperti titik (.), koma (,), spasi dan tanda pemisah lain menjadi kata saja, baik berupa kata penting maupun kata tidak penting. Token yang digunakan dalam penelitian ini menggunakan uni-gram, dengan mengimplementasikan tokenisasi uni-gram sebagai fitur kata untuk memuat kata-kata yang ada di sebuah kalimat. Contoh dari tokenizing dapat dilihat pada gambar 2.2.



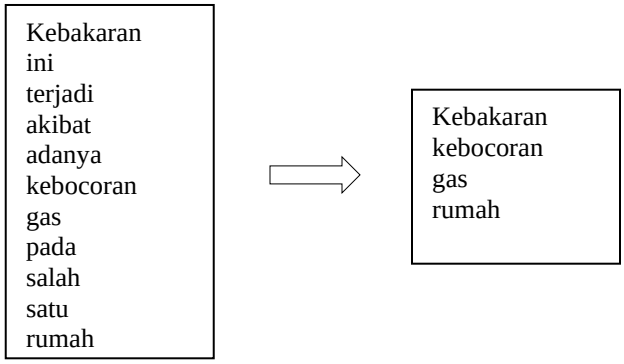
Gambar 2.2 Contoh Tokenizing

Komponen khas yang biasa ada pada tweet seperti username, URL, “RT” (tanda retweet), karena tanda khas komponen tersebut tidak memiliki pengaruh terhadap nilai sentimen, maka komponen tersebut dihilangkan. Komponen username diidentifikasi dengan kemunculan karakter “@”, sedangkan komponen URL dikenali melalui ekspresi regular (http). Contoh dari normalisasi fitur dapat dilihat pada gambar 2.3.



Gambar 2.3 Contoh Normalisasi Fitur

Kata yang terlalu sering muncul pada setiap dokumen tidak terlalu baik digunakan sebagai kata kunci. Faktanya, kata yang muncul sampai 80% pada setiap dokumen tidak berguna untuk pengambilan informasi. Kata-kata yang sering muncul tersebut dikenal sebagai stopword. Stopword adalah kata-kata yang sering muncul dalam jumlah besar dan tidak memiliki makna. Contoh dari stopword untuk bahasa indonesia yaitu adalah, adanya, akhirnya, bahkan, bagaimanapun. Dengan dibuangnya stopword, ukuran kosakata menjadi berkurang sehingga hanya kata-kata penting yang terdapat dalam dokumen dan diharapkan akan menjadi kata-kata yang memiliki bobot yang tinggi. Contoh stopword removal dapat dilihat pada gambar 2.4.



Gambar 2.4 Contoh Stopword Removal

Emoticon (emotion icon) merupakan salah satu cara pengungkapan perasaaan secara tekstual. Hal ini tentu akan membantu dalam menentukan sentimen dari suatu kalimat. Setiap emoticon akan dikonversikan ke dalam string yang sesuai. Contoh convert emoticon dapat dilihat pada tabel I.

TABEL I
CONVERT EMOTICON

Emoticon	Konversi
:D :-D =D XD	Eketawa
:) :-) 8) =) ^_^ ^^	Esenang
:v :-v :-/ :/ :-/ =/ -_-	Ekesal
:(:-(:'(;(>.< ><	Esedih

Kata-kata yang sering muncul dalam tweet cenderung tidak baku, menggunakan kata gaul dan tidak ada dalam kamus seperti gue, loe dan lain-lain, serta tidak jarang yang menggunakan potongan kata misalnya yg, brp, bgm, smoga dan lain sebagainya. Oleh karenanya untuk kata yang tidak ada dalam kamus akan di terjemahkan ke kata terdekat dengan menggunakan kamus yang dibuat untuk melihat pola kemunculan kata-kata tidak baku tersebut. Contoh dari normalisasi kata dapat dilihat pada tabel II.

TABEL II
NORMALISASI KATA

Kata	Kata Ganti
Yg	yang
brp	berapa
Bgm	bagaimana
smoga	semoga
Gue	saya

Stemming adalah suatu proses yang menyediakan suatu pemetaan antara berbagai kata dengan morfologi yang berbeda menjadi satu bentuk dasar (stem) [3]. Stemming kata Indonesia menghadapi persoalan variasi imbuhan (*affix*) yang lebih kompleks dibandingkan dengan bahasa inggris [4]. Sebagai contoh untuk menghilangkan imbuhan (affix) yang dapat berupa awalan (*prefixes*), akhiran (*suffixes*) dan sisipan (*infixes*) atau kombinasi (*confixes*) untuk memperoleh akar kata harus dilakukan dengan pertimbangan yang rumit menyangkut urutannya. Untuk kasus sederhana seperti kata "minuman" memiliki kata dasar "minum" dan akhiran "an", persoalan dapat dianalogikan seperti penghilangan akhiran "s" atau akhiran "ed" dalam bahasa inggris. Penghilangan imbuhan untuk beberapa kasus berikut tidak dijumpai analoginya dalam bahasa inggris:

a) "pemerintah", diturunkan dari kata dasar "perintah" mendapat sisipan "em".
b) "buku-buku", bentuk jamak dari buku.
c) "pemberdayaannyapun" memiliki kata dasar "daya" mendapat awalan "pe" dan "ber-" dan akhiran "an", "nya", dan "-pun".

Dalam bahasa Indonesia imbuhan dapat dikelompokkan menjadi beberapa kelompok [5], yaitu:

1. Akhiran infleksi (*inflection suffixes*), yaitu akhiran yang tidak mengubah akar kata, misalnya "duduk" dapat ditambah akhiran "-lah" menjadi "duduklah". Infleksi dapat dibagi dua yaitu:
   a) Partikel, yaitu: "-lah", "kah", ataupun "-pun"
   b) Kata ganti milik (possesive pronoun), yaitu: "-ku", "-mu", "-nya"
   Partikel dan possesive pronoun dapat muncul bersamaan dan bila terjadi maka possesive pronoun akan mendahului partikel, misalnya dalam kata "rumahnyapun" atau "mobilmulah".
2. Akhiran derivasi (*derivation suffixes*) yaitu akhiran yang diterapkan langsung pada akar kata membentuk kata baru. Hanya ada satu akhiran derivasi untuk satu kata. Kata "lapor" dapat ditambah dengan akhiran derivasi "-kan" menjadi "laporkan" dan ditambah dengan akhiran infleksi "lah" menjadi "laporkanlah".
3. Awalan derivasi (*derivation prefixes*) awalan yang dapat diterapkan langsung pada akar kata, misalnya "perawat" atau dapat diterapkan pada kata yang sudah berawalan seperti kata "pemberangkatan".

Algoritma Arifin dan Setiono sebagai algoritma stemming untuk teks berbahasa indonesia, walaupun menggunakan kamus seperti algoritma Nazief dan Adriani mengajukan proses stemming yang lebih sederhana[1]. Pada algoritma Arifin dan Setiono menghilangkan seluruh kemungkinan akhiran. Setiap kali awalan atau akhiran dihilangkan program mengecek ke dalam kamus.

Algoritma Arifin dan Setiono mengasumsikan bahwa setiap kata memiliki dua awalan dan tiga akhiran yaitu:

[AW1] + [AW2] + KD + [AK3] + [AK2] + [AK1]

Dimana AW = awalan, KD = kata dasar dan AK = akhiran
Langkah-langkah stemming algoritma Arifin dan Setiono yaitu:

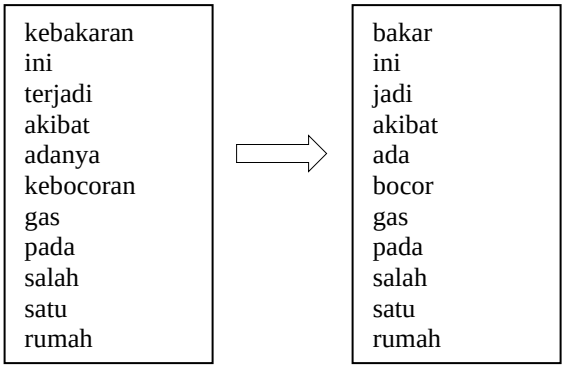
1. Lakukan pemeriksaan setiap kata, siapkan variabel p1,p2,s1,s2,s3
2. Pemotongan dilakukan secara berurut yaitu:
   a) Awalan I, hasil disimpan pada p1
   b) Awalan II, hasil disimpan pada p2
   c) Akhiran I, hasil disimpan pada s1
   d) Akhiran II, hasil disimpan pada s2
   e) Akhiran II, hasil disimpan pada s3

Setiap tahap pemotongan hasil dicek dalam kamus, jika ada dalam kamus algoritma selesai jika tidak ada proses dilanjutkan ke pemotongan berikutnya.

3. Jika sampai pada langkah 2.e. belum ditemukan dalam kamus maka dilakukan proses kombinasi. Kata dasar yang dihasilkan dikombinasikan dengan imbuhan-imbuhan dalam 12 kombinasi yaitu:
   a) Kata Dasar
   b) Kata Dasar + AK III
   c) Kata Dasar + AK III + AK II
   d) Kata Dasar + AK III + AK II + AK I
   e) AW I + AW II + Kata Dasar
   f) AW I + AW II + Kata Dasar + AK III
   g) AW I + AW II + Kata Dasar + AK III + AK II

h) AW I + AW II + Kata Dasar + AK III + AK II + AK I
i) AW II + Kata Dasar
j) AW II + Kata Dasar + AK III
k) AW II + Kata Dasar + AK III + AK II
l) AW II + Kata Dasar + AK III + AK II + AK I

Kelebihan algoritma ini dibandingkan dengan algoritma Nazief dan Adriani adalah dalam hal mengatasi *overstemming* yaitu jika sebagian kata dasar di stem karena sebagai awalan atau akhiran. Sebagai contoh kata "diselamatkan" memiliki akar kata "selamat". Langkah pertama awalan I "di" dibuang kata yang dihasilkan adalah "selamatkan", karena kata tidak ada dalam kamus proses dilanjutkan dengan menghilangkan awalan II "se" menghasilkan "lamatkan" yang merupakan kekeliruan karena "se" bukan awalan tetapi bagian dari akar kata. Setelah penghilangan awalan selesai proses dilanjutkan dengan penghilangan akhiran "-kan" menghasilkan "lamat", karena "lamat" tidak ada dalam kamus maka dilanjutkan pada langkah 3 dengan mencoba menggabungkan kembali berbagai kombinasi imbuhan dan setelah menggabungkan awalan "se" menghasilkan "selamat" ternyata ada dalam kamus maka proses stemming berhasil. Contoh pada gambar 2.5.



Gambar 2.5 Contoh Stemming

Analisis Sentimen atau dapat disebut sebagai opinion mining adalah bidang studi dalam menganalisis pendapat orang, evaluasi, penilaian, sikap dan emosi terhadap suatu entitas seperti produk, jasa, organisasi, individu, isu-isu, peristiwa dan topik. Fokus utama dalam analisis sentimen adalah untuk menyatakan hal yang termasuk opini positif dan hal yang termasuk opini negatif. Salah satu contoh penggunaan analisis sentimen dalam kehidupan nyata adalah identifikasi kecenderungan pasar dan opini pasar terhadap suatu objek produk, selain mengekstraksi sentimen, hal yang biasa orang ingin ketahui adalah kapan terjadi perubahan sentimen dan penyebab perubahan sentimen tersebut. Hal ini menjadi penting, karena dengan mengetahui penyebab sentimen berubah, pihak yang bersangkutan bisa mengambil keputusan dengan lebih baik. Contohnya ketika mengetahui suatu topik atau kejadian menyebabkan sentimen turun, maka pihak yang bersangkutan akan menghindari kejadian tersebut untuk meningkatkan sentimen. Kebutuhan-kebutuhan tersebut biasanya muncul ketika suatu pihak ingin mendapatkan sentimen publik yang baik atau dengan pencitraan, kebutuhan seperti ini biasa dimiliki oleh layanan provider telepon.

Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data. Tujuan dari klasifikasi adalah untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya.

Teorema Bayes adalah pendekatan statistik yang fundamental dalam pengenalan pola (*pattern recognation*). Metode bayes juga merupakan metode yang baik di dalam machine learning berdasarkan data training, dengan menggunakan probabilitas bersyarat sebagai dasarnya.

Pada teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan X dan Y), maka teorema Bayes dirumuskan sebagai berikut:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Keterangan:
X = data sampel dengan kelas (label) yang tidak diketahui.
Y = hipotesa bahwa X adalah data dengan kelas (kelas yang sudah diketahui).
P(Y) = peluang dari hipotesa Y.
P(X) = peluang data sampel yang diamati.
P(X|Y) = peluang data sampel X, bila diasumsikan bahwa hipotesa benar.

Naive Bayesian Classification (NBC) merupakan sebuah metode klasifikasi berakar pada teorema Bayes. Ciri utama dari Naive Bayesian Classification adalah asumsi yang sangat kuat akan independensi dari masing-masing variabel, dengan kata lain Naive Bayesian Classification mengasumsikan bahwa keberadaan sebuah variabel tidak ada kaitannya dengan keberadaan variabel yang lain. Algortima Naive Bayesian Classification terdiri dari dua tahapan. Tahap pertama adalah pelatihan terhadap himpunan dokumen, contoh data latih dan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya (kelas).

Algoritma ini memanfaatkan teori probabilitas yang dikemukakan oleh ilmuwan inggris Thomas Bayes yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman masa sebelumnya, karena asumsi atribut tidak saling terkait, maka:

$$Vmap = \underset{Vj \in V}{argmax}\ P(Vj) \prod P(w_{k} | Vj)$$

Setelah diperoleh perhitungan untuk masing-masing kategori, maka kategori yang dipilih adalah yang memiliki nilai Vmap terbesar. Nilai P(Vj) ditentukan pada saat pelatihan, yang nilainya berdasarkan persamaan:

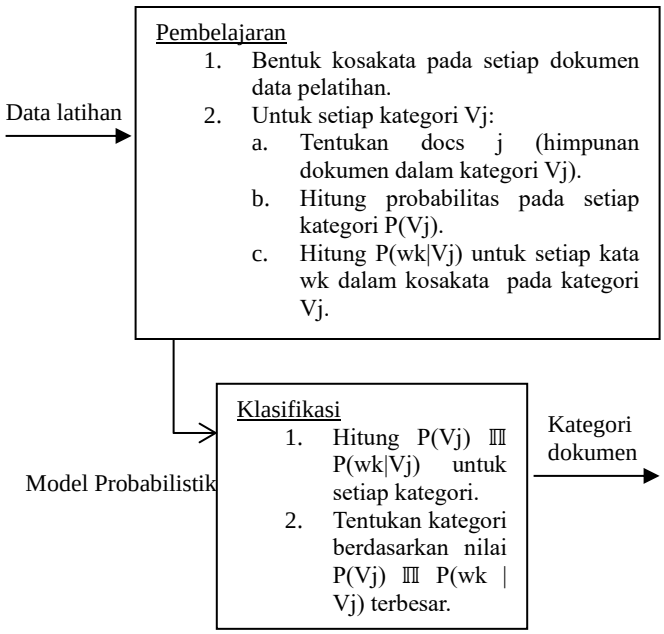
$$P(Vj) = \frac{|docs\ j|}{|contoh|}$$

Keterangan:
P(Vj) = probabilitas setiap dokumen terhadap sekumpulan dokumen.
|docs j| = banyak dokumen yang memiliki kategori j dalam pelatihan.
|contoh| = banyaknya dokumen dalam contoh yang digunakan saat pelatihan.

Untuk nilai $P(w_k \mid V_j)$ ditentukan dengan persamaan:

$$P(w_k \mid V_j) = \frac{|nk+1|}{n+|kosakata|}$$

Keterangan:
$P(w_k \mid Vj)$ = probabilitas kemunculan kata wk pada suatu dokumen dengan kategori Vj.
nk = frekuensi munculnya kata wk dalam dokumen yang berkategori Vj.
n = banyaknya seluruh kata dalam dokumen berkategori Vj.
|kosakata| = banyaknya kata dalam contoh pelatihan.



Gambar 2.6 Tahapan Algoritma Naive Bayesian Classification

III. ANALISA DAN DESAIN SISTEM

A. Penerapan Metode

Data yang digunakan dalam penelitian ini diambil dari kumpulan tweets bahasa indonesia yang diambil dari official akun twitter Telkomsel, Indosat, serta XL Axiata. Data tweets ini diperoleh secara manual dari Twitter API tanpa menggunakan program crawler. Data tweets yang diambil mengandung kata "telkomsel", "kartu_as", "simpati", "indosatmania", "indosat", "indosatcare", "xl123", "xlcare", "xlandme".

Penulis mengikuti teknik yang dilakukan oleh peneliti terdahulu dengan menggunakan kata-kata bermakna sentimen sebagai penanda pada tweets tersebut. Penulis menggunakan sebuah kata yang memiliki sentimen sebagai kata kunci. Berikut adalah daftar kata-kata yang digunakan sebagai kata kunci pada tabel III.

TABEL III
DATA AWAL

Jenis Kata	Kata
Kata positif	terimakasih, baik, betul, sukses, pasti, sip, top, baik, bagus, lancar, senang, beruntung, bahagia, selamat, asik, puas
Kata negative	lambat, lemot, lelet, payah, bodoh, kecewa, parah, gagal, sedih, susah, lemah, rusak, mahal
Kata netral	telkomsel, kartu, perdana, undian,paket,indosat,xl, hadiah

Data tweets yang terkumpul akan melalui tahap preprocessing dan selanjutnya akan diklasifikasikan. Dalam sistem analisis sentimen, tweets akan diklasifikasikan ke dalam tiga kelas, yaitu kelas sentimen positif, kelas sentimen negatif dan kelas sentimen netral. Contoh dari tweets yang termasuk sentimen positif dapat dilihat pada gambar 3.1, tweets yang termasuk sentimen negatif dapat dilihat pada gambar 3.2, tweets yang termasuk sentimen netral dapat dilihat pada gambar 3.3.

Gambar 3.1 Contoh Sentimen Positif

Gambar 3.2 Contoh Sentimen Negatif

Data yang dibutuhkan dalam penelitian ini terdiri dari dua jenis, yaitu data latih dan data uji. Data latih yang digunakan ini diambil dari kumpulan tweets yang telah dilabeli dengan kelas sentimennya secara manual. Data inilah yang digunakan sebagai data latih untuk membentuk model analisis sentimen. Model ini nantinya akan digunakan untuk mengklasifikasikan tweets pada kelas sentimennya. Pada penelitian ini, metode klasifikasi yang digunakan adalah naïve bayes classifier. Data uji ini menggunakan kumpulan tweets yang belum memiliki label.

Gambar 3.3 Contoh Sentimen Netral

Pemrosesan text merupakan proses menggali, mengolah, mengatur informasi dengan cara menganalisis hubungannya aturan-aturan yang ada di data texstual semi terstruktur atau tidak terstruktur. Preprocessing merupakan salah satu langkah yang penting dalam analisis sentimen. Sama halnya preprocessing pada Information Retrieval (IR), tahapannya terdiri dari case folding, tokenizing, normalisasi fitur, stopword removal, stemming. Namun pada preprocessing analisis sentimen ada tambahan tahapan seperti convert emoticon.

Tahapan ini semua huruf akan diubah menjadi huruf kecil. Langkah-langkah case folding dalam contoh salah satu tweets: "Terimakasih ya min, kartunya sudah terbuka :D @Telkomsel". Bila ditemukan karakter yang mengandung huruf kapital maka huruf tersebut akan diubah menjadi huruf kecil.

Gambaran tahap case folding dapat dilihat pada tabel 3.1.

Tabel 3.1 Contoh Case Folding

Terimakasih ya min, kartunya sudah terbuka :) @Telkomsel
⇩
terimakasih ya min, kartunya sudah terbuka :) @telkomsel

Twitter biasa memiliki komponen khas seperti username, URL (*Uniform Resource Locator*) dan "RT (tanda retweet)". Karena username, URL, "RT" tidak berpengaruh pada nilai sentimen maka komponen tersebut akan dibuang. Komponen username diwakili dengan kemunculan karakter "@". Selain username, karakter "@" digunakan juga untuk menunjukkan suatu tempat seperti @kickfest. Nama tempat tersebut tidak berpengaruh pada analisis sentimen sehingga nama tempat harus dihapus. Pada komponen URL ditunjukkan dengan "www" dan "http". Tahapan normalisasi fitur menggunakan kata dari hasil case folding, kata dari case folding di periksa bila menemukan username, URL dan RT maka akan dihilangkan.

Gambaran tahap normalisasi fitur dapat dilihat pada tabel 3.2.

Tabel 3.2 Contoh Normalisasi Fitur

terimakasih ya min, kartunya sudah terbuka :) @telkomsel
⇩
terimakasih ya min, kartunya sudah terbuka :) telkomsel

Tahap ini, kumpulan tweets yang terdapat emoticon akan dikonversi ke dalam string yang sesuai. Namun tidak semua akan diimplementasikan, karena tidak semua emoticon banyak digunakan oleh pengguna Twitter. Emoticon yang digunakan dapat dilihat pada table I. Langkah-langkah dalam tahap convert emoticon menggunakan kata yang berasal dari normalisasi fitur dilanjutkan dengan membandingkan setiap karakter dengan list emoticon dan bila terdapat emoticon maka langsung diubah ke dalam bentuk string. Gambaran tahap convert emoticon dapat dilihat pada tabel 3.3.

Tabel 3.3 Contoh Convert Emoticon

terimakasih ya min, kartunya sudah terbuka :) telkomsel
⇩
terimakasih ya min, kartunya sudah terbuka eketawa telkomsel

Tahap ini akan dilakukan pengecekan tweets dari karakter pertama sampai dengan karakter terakhir. Apabila karakter ke-i bukan tanda pemisah kata seperti titik(.), koma(,), spasi dan tanda pemisah lainnya maka akan digabungkan dengan karakter selanjutnya. Langkah-langkah tahap tokenizing menggunakan kata dari hasil convert emoticon dilanjutkan memotong setiap kata dalam teks berdasarkan pemisah kata seperti titik(.), koma(,) dan spasi. Bagian yang termasuk dalam daftar emoticon tidak dibuang. Bagian yang hanya memiliki satu karakter non alfabet dan angka akan dibuang. Dengan fitur uni-gram membuat setiap kata yang terseleksi menjadi potongan karakter.

Gambaran tahap tokenizing dapat dilihat pada tabel 3.4.

Tabel 3.4 Contoh Tokenizing

terimakasih ya min, kartunya sudah terbuka eketawa telkomsel
⇩

terimakasih	Ya	min	kartunya
Sudah	Terbuka	eketawa	telkomsel

Tahap ini, tweets yang telah melalui tahap tokenizing akan melalui tahap normalisasi kata untuk memperbaiki kata-kata yang tidak baku biasanya sering muncul dalam tweet. Gambaran tahap tokenizing dapat dilihat pada tabel 3.5.

Tabel 3.5 Contoh Normalisasi Kata

terimakasih	ya	min	kartunya
sudah	terbuka	eketawa	telkomsel

⇩

terimakasih	Iya	admin	kartunya
sudah	terbuka	eketawa	telkomsel

Tahap ini, kumpulan tweets yang telah melewati tahap normalisasi kata akan melalui tahap stopword removal. Setiap kata pada tweets akan diperiksa bila terdapat kata yang tidak ada hubungan dalam analisis sentimen maka akan dihilangkan. Langkah-langkah tahap stopword removal dengan membandingkan hasil normalisasi kata dengan daftar stopword dilanjutkan pengecekan kata dengan daftar stopword, bila ada kata yang terdapat pada daftar stopword maka akan dihilangkan.

Kata-kata yang muncul dalam tweets banyak variasi. Sehingga setiap kata-kata tersebut diambil bentuk kata dasarnya dengan cara menghilangkan awalan atau akhiran. Langkah-langkah pada tahap stemming menggunakan kata dari hasil stopword removal dilanjutkan setiap kata dalam tweets diperiksa dari awal sampai akhir kata jika terdapat imbuhan maka imbuhan tersebut akan dihilangkan.

Gambaran tahap stemming dapat dilihat pada tabel 3.6. Proses training naive bayesian classification bertujuan mencari keyword beserta probabilitasnya yang akan digunakan pada proses testing. Pada proses ini di bagi menjadi dua kelas, yaitu:

1. Data 1 (D1) = kelas sentimen positif
2. Data 2 (D2) = kelas sentimen negatif
3. Data 3 (D3) = kelas sentimen netral

Tabel 3.6 Contoh Stemming

terima	Kasih	admin	Kartunya
eketawa	Telkomsel		

terima	kasih	admin	kartu
eketawa	telkomsel		

Contoh tabel data training dapat dilihat pada tabel IV

TABEL IV
CONTOH DATA TRAINING

Data	Keyword(kemunculan)	Kelas Sentimen
D1	Bagus(8), Cepat(2), Ramah(2)	Positif
D2	Kecewa(3), Lambat(4), Rugi(4)	Negatif
D3	Telkomsel(10), Kartu(3), Paket(5)	Netral

Keyword yang dihasilkan sebanyak 9 kata. Selanjutnya dilanjutkan dengan mencari nilai probabilitas pada keywordnya menggunakan persamaan:

$$P(Vj) = \frac{|docs\ j|}{|contoh|}$$

Keterangan:
P(Vj) = probabilitas setiap dokumen terhadap sekumpulan dokumen.
|docs j| = banyak dokumen yang memiliki kategori j dalam pelatihan.
|contoh| = banyaknya dokumen dalam contoh yang digunakan saat pelatihan.

Jika dokumen-dokumen teks tersebut dikelompokkan (Classification) ke dalam dua kelas, C1 (Pemograman) dan C5 (Umum) maka kita dapat memperoleh hasil sebagai berikut:
C1 = akan beranggotakan D1, D2
C5 = akan beranggotakan D3

TABEL V
NILAI P(VJ) PADA SETIAP KELAS

Data	Keyword(kemunculan)	Kelas Sentimen	P(Vj)
D1	Bagus(8), Cepat(2), Ramah(2)	Positif	1/3 = 0.3
D2	Kecewa(3), Lambat(4), Rugi(4)	Negatif	1/3 = 0.3
D3	Telkomsel(10), Kartu(3), Paket(5)	Netral	1/3= 0.3

$$P(\text{positif}) = \frac{|Positif|}{(Positif|Negatif|Netral)} = \frac{1}{3} = 0.3$$

$$P(\text{negatif}) = \frac{|Negatif|}{(Positif|Negatif|Netral)} = \frac{1}{3} = 0.3$$

$$P(\text{netral}) = \frac{|Netral|}{(Positif|Negatif|Netral)} = \frac{1}{3} = 0.3$$

Untuk perhitungan setiap kata Wk pada kelas Vj menggunakan persamaan:

$$P(w_k \mid V_j) = \frac{|nk+1|}{n+|kosakata|}$$

Keterangan:
$P(w_k \mid Vj)$ = probabilitas kemunculan kata wk pada suatu dokumen dengan kategori Vj.
n_k = frekuensi munculnya kata wk dalam dokumen yang berkategori Vj.
n = banyaknya seluruh kata dalam dokumen berkategori Vj.
|kosakata| = banyaknya kata dalam contoh pelatihan.

Proses testing ini, data uji akan melewati proses klasifikasi berdasarkan data training. Data testing adalah data tweets yang belum diklasifikasikan. Proses pemilihan data Testing dapat dilakukan secara manual dan random. Pemilihan secara random dilakukan dengan menginputkan jumlah data yang akan di uji, sedangkan pemilihan secara manual dilakukan dengan memilih satu per satu tweet untuk analisis sentimen. Data testing adalah hasil dari tahap preprocessing dapat dilihat pada tabel 3.8.

Tabel 3.8 Hasil Preprocessing

bagus	telkomsel	Cepat	ramah

Data testing (D4): bagus(1), telkomsel(1), cepat(1), ramah(1).

Untuk perhitungan Vmap menggunakan persamaan:

$$Vmap = \underset{Vj \in V}{argmax}\ P(Vj) \prod P(w_k \mid Vj)$$

Berdasarkan persamaan di atas diperoleh perhitungan sebagai berikut:

$$Vmap = \underset{Vj\{\text{positif,negatif,"}netral\}}{argmax} P(Vj)\ P(\text{“bagus”} \mid Vj)\ P(\text{“telkomsel”} \mid Vj)\ P(\text{“cepat”} \mid Vj)\ P(\text{“ramah”} \mid Vj)$$

Nilai Vmap untuk Sentimen positif

Vmap(“Positif”) = P(“Positif”) P(“bagus” | “Positif”)
P(“telkomsel” | “Positif”)
P(“cepat” | “Positif”) P(“ramah” | “Positif”)
= 1/3 x 9/21 x 1/21 x 3/21 x 3/21
= 1,3E-04

Nilai Vmap untuk sentimen negatif

Vmap(“Negatif”) = P(“Negatif”) P(“bagus” | “Negatif”)
P(“telkomsel” | “Negatif”)
P(“cepat” | “Negatif”) P(“ramah” | “Negatif”)
= 1/3 x 1/20 x 1/20 x 1/20 x 1/20
= 2,0E-06

Nilai Vmap untuk sentimen netral

Vmap(“Netral”) = P(“Netral”) P(“bagus” | “Netral”)
P(“telkomsel” | “Netral”) P(“cepat” | “Netral”) P(“ramah” | “Netral”)
= 1/3 x 1/27 x 11/27 x 1/27 x 1/27
= 6,9E-06

Berdasarkan perhitungan diatas, didapatkan nilai Vmap tertinggi untuk kelas sentimen positif di bandingkan dengan

kelas sentimen negatif dan sentimen netral sehingga tweets tersebut termasuk kelas sentimen positif. Peluang kemunculan kata yang besar akan menghasilkan Vmap yang tinggi, sehingga dokumen data uji akan terklasifikasi ke dalam karakter dengan Vmap yang paling tinggi. Jika nilai Vmap untuk kelas sentimen positif dan negatif sama, maka tweets tersebut termasuk dalam sentimen negatif karena dengan demikian perusahaan akan mengkaji ulang kekurangan produknya.

Untuk menghitung akurasi dari hasil klasifikasi maka diperlukan data hasil klasifikasi dengan keseluruhan data yang sebenarnya. Berikut perhitungan naïve bayes:

1. True positif = 90
2. True negatif = 95
3. True netral = 52
4. False positif = 17
5. False negatif = 8
6. False netral = 38

Jumlah data uji yang benar = 237
Jumlah data uji yang salah = 63
Total data uji = 300

$$\text{Akurasi} = \frac{90+95+52}{90+95+52+17+8+38} = \frac{237}{300} \text{ x } 100\% = 79\ \%$$

IV. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan pada sistem analisis sentimen dengan menggunakan metode naïve bayes classification ini maka dapat diambil kesimpulan sebagai berikut:

1. Aplikasi ini mampu melakukan pengklasifikasian sentimen secara acak maupun manual.
2. Proses pengklasifikasian semakin akurat dengan menggunakan metode confusion matric sebagai proses perhitungan akurasi.
3. Metode naïve bayes classification mampu mengklasifikasikan tweet pada sistem analisis sentimen

REFERENSI

[1] Ronen Feldman, James Sanger. 2007. *The Text Mining Handbook : Advanced Approaches In Analyzing Unstructured Data*. Cambridge University Press : England.

[2] Lailil Muflikhah, Yugo Yudansha, MT Marji. 2013. *Sentiment Analysis pada Review Barang Berbahasa Indonesia dengan Metode K-NN*. 1-7.

[3] Fadillah Z Tala. 2003. *A Study of Stemming Efects on Information Retrieval in Bahasa Indonesia*. Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands.

[4] Albert Bifet, Carlos Castillo, Paul-Alexandru Chirita and Ingmar Weber. 2005. *An Analysis of Factors Used in Search Engine Ranking*. airweb.cse.lehigh.edu/2005/bifet.pdf

[5] Bobby Nazief dan Mirna Adriani. 1996. *Confixstripping : Approach to Stemming Algorithm for Bahasa Indonesia*. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta.

Penulis I, Ronny Julianto**,** memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi STMIK PPKIA Tarakanita Rahmawati Tarakan, lulus tahun 2016..

Penulis II, Evi Dianti Bintari**,** memperoleh gelar Magister Komputer (M.Kom), Sekolah Tinggi Teknik Surabaya. Saat ini menjadi Dosen di STMIK PPKIA Tarakanita Rahmawati

Penulis III, Indrianti**,** memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi STMIK PPKIA Tarakanita Rahmawati Tarakan, lulus tahun 2012. Saat ini menjadi Dosen di STMIK PPKIA Tarakanita Rahmawati