IMPLEMENTASI LEXICON BASED DAN NAIVE BAYES PADA ANALISIS SENTIMEN PENGGUNA TWITTER TOPIK PEMILIHAN PRESIDEN 2019

[1]Gusti Nur Aulia, [2]Eka Patriya
[1,2]Fakultas Teknologi Industri Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
[2]ekapatriya@staff.gunadarma.ac.id,

Abstrak

Pilpres saat ini cukup menyita perhatian, karena berbagai rumor yang beredar. Masyarakat juga menjadi sasaran elit politik, dimana suara mereka merupakan penentu keberlangsungan arah politik untuk lima tahun kedepan. Opini-opini positif, netral maupun negatif dapat menimbulkan ancaman munculnya berita bohong (hoax). Salah satu sarana yang digunakan masyarakat dalam mengekspresikan pilihan politiknya adalah melalui media sosial salah satunya twitter. Data seperti opini publik dapat diolah menjadi sebuah informasi yang bermanfaat, salah satunya melalui analisis sentimen. Pada penelitian ini, akan dilakukan analisis sentimen pada Twitter tentang pemilihan presiden 2019. Tahapan analisis sentimen pada penelitian ini terdiri dari akuisisi data, pre-processing, klasifikasi data, evaluasi data dan visualisasi data. Preprocessing dilakukan dengan case folding, normalisasi data, filtering, ubah kata baku, stopword dan stemming. Penelitian ini melakukan 2 metode yaitu dengan metode Lexicon Based dan Naïve Bayes Classifier. Hasil akhir dari analisis kemudian dihitung nilai akurasi menggunakan confusion matrix dan di visualisasikan menggunakan web server. Penentuan sentimen prediksi dilakukan menggunakan metode Lexicon Based dan Labelisasi dengan perhitungan secara manual. Data latih dan data uji akan digunakan dalam proses pelatihan dan pengujian menggunakan Naive Bayes Classifier. Hasil klasifikasi yang dilakukan oleh metode Naive Bayes Classifier disebut sentimen aktual. Perhitungan tingkat keakurasian antara sentimen prediksi terhadap sentimen aktual menggunakan pengujian confusion matrix. Hasil yang didapatkan adalah tingkat akurasi antara sentimen prediksi dan sentimen aktual dengan Lexicon Based sebesar 64,49% pada data uji dan pada data latih sebanyak 94,2% serta dengan menggunakan Labelisasi dan Naive Bayes Classifier sebesar 86,53% pada data uji dan data latih sebesar 94,08%. Hasil penelitian ini diharapkan dapat membantu melakukan riset atas opini masyarakat pada Twitter mengenai Pilpres 2019 yang mengandung sentimen positif, negatif atau netral.

***Kata Kunci**: Classifier, Lexicon Based, prediksi, Naive Bayes,* Twitter

Abstract

In this study, researchers will conduct a sentiment analysis on Twitter about the election presidential 2019. The stages of sentiment analysis in this study consist of data acquisition, pre-processing, data processing or classification, data evaluation and data visualization. The pre-processing processes are case folding, data normalization, filtering, changing standard words, stopword and stemming. This study conducted 2 methods, namely the Lexicon Based method and the Naïve Bayes Classifier. The final result of the analysis is then calculated the accuracy value using a confusion matrix and visualized using a web server. Determination of initial sentiments or commonly referred to as predictive sentiments is done using the Lexicon Based method and Labeling with manual calculations. After determining the initial sentiment, the data is divided into training data and test data. In this study the number of training data is 845 data (data to 1-845) and for test data amounted to 245 data (data to 846-1090). Training data and test data will be used in the training and testing process using Naive Bayes Classification.

The results of the classification carried out by the Naive Bayes Classification method are called actual sentiments. After the prediction of sentiment is determined at the beginning of the process and the actual sentiment has been obtained, then the level of accuracy is calculated between the predicted sentiment to the actual sentiment using the confusion matrix test. The results obtained are the level of accuracy between predicted sentiments and actual sentiments with Lexicon Based of 64,49% in the test data and in training data as much as 94,2% and by using Labeling and Naive Bayes Classification of 86,53% in the test data and data training of 94,08%.

***Keywords** : Classifier, Lexicon Based, prediction, Naive Bayes, Twitter*

PENDAHULUAN

Indonesia adalah salah satu negara yang menganut sistem demokrasi, sebagai negara yang menganut sistem demokrasi, penting bagi warga negara Indonesia memiliki sebuah proses untuk memilih orang yang dapat mengisi jabatan-jabatan politik tertentu. Proses tersebut kita kenal sebagai Pemilu atau Pemilihan Umum. Di Indonesia sendiri diketahui bahwa pemilu diselenggarakan secara periodik yaitu selama lima tahun sekali. Pada tanggal 17 April 2019 yang lalu telah diselenggarakan kembali pemilu, salah satunya adalah pemilihan presiden (Pilpres). Pilpres saat ini cukup menyita perhatian, karena berbagai rumor beredar saling menjatuhkan antar kedua pasangan calon presiden (paslon). Masyarakat juga menjadi sasaran elit politik, dimana suara mereka merupakan penentu keberlangsungan arah politik untuk lima tahun kedepan. Opini-opini positif, netral maupun negatif seperti SARA, Hak Asasi Manusia (HAM) dan ekonomi dapat menimbulkan ancaman munculnya berita bohong (*hoax*). Salah satu sarana yang digunakan masyarakat dalam mengekspresikan pilihan politiknya adalah melalui media sosial salah satunya twitter. Twitter merupakan jejaring sosial yang populer di kalangan pengguna internet saat ini, karena menyediakan banyak fitur yang menarik untuk digunakan oleh pengguna seperti berita, tweet antar sesama pengguna dan lain sebagainya. Data seperti opini publik dapat dikumpulkan dan diolah menjadi sebuah informasi yang bermanfaat, salah satunya melalui analisis sentimen [1]. Analisis sentimen diimplementasikan untuk mengklasifisikan data [2] kedalam klasifikasi yang bersifat positif, negatif maupun netral.

Penelitian terkait mengenai analisis sentimen pada Twitter dilakukan peneliti terdahulu. Penelitian dilakukan [3] dalam melakukan analisis sentimen Twitter menggunakan metode *Lexicon Based* dan *Double Propagation*. Kombinasi *Lexicon Based* dan *Double Propagation* mampu menghasilkan 7 parameter analisis sentimen yaitu sangat positif, positif, agak positif, netral, agak negatif, negatif dan sangat negatif. Penelitian [4] membuat aplikasi klasifikasi opini yang menerapkan pendekatan *Naive Bayes* untuk mengklasifikasikan kata-kata dan difokuskan pada *tweets* dalam Bahasa Indonesia. Aplikasi ini kemudian diterapkan untuk mengklasifika-

sikan opini publik pada Twitter terkait layanan pemerintah terhadap masyarakat, berdasarkan sentimen positif, negatif atau netral. Data latih diperoleh melalui aplikasi platform KNIME *Analytic* dan sumber teks diperoleh dari akun Twitter Dinas pemerintah Kota Bandung. Penelitian [5] penelitian ini mencoba menganalisis persepsi masyarakat kedalam kelas sentimen menggunakan metode *Lexicon Based* dengan *SentiWordNet*. *Dataset* yang digunakan adalah *tweets* mengenai kenaikan harga rokok dalam Bahasa Indonesia berjumlah 350 buah. Data diklasifikasikan sesuai SentiWordNet pada tiap-tiap kata dalam kalimat. Untuk kata yang memiliki lebih dari satu arti maka *synset* dipilih berdasarkan metode *First Sense* dari *SentiWordNet* yang muncul paling popular. Peneliti (Antinasari, Perdana, & Fauzi, 2017), pada penelitian ini digunakan kamus kata tidak baku dan normalisasi *Levenshtein Distance* untuk memperbaiki kata yang tidak baku menjadi kata baku dengan peng-klasifikasian *Naïve Bayes*. Peneliti [6] melakukan penelitian analisis sentimen, untuk *preprocessing* data menggunakan *tokenisasi*, *cleansing* dan *filtering*, untuk menentukan *class* sentimen dengan metode *Lexicon Based*. Untuk proses klasifikasinya menggunakan metode *Naïve Bayes Classifier (NBC)* dan *Support Vector Machine* (SVM). Peneliti [7] melakukan penelitian yang terdahulu tentang ujaran kebencian. Metode yang digunakan dalam mengolah data dokumen tersebut adalah *Backpropagation Neural Network* dengan pembaruan fitur menggunakan *Lexicon Based Features* yang dikombinasikan dengan *Bag of Words*.

Pada penelitian ini, peneliti akan mengimplementasikan penggunaan metode *Lexicon Based* dan *Naïve Bayes Classifier* pada penentuan klasifikasi sentimen positif, negatif maupun netral dalam analisis sentimen mengenai pemilihan presiden 2019. Hasil klasifikasi tersebut kemudian akan di hitung akurasinya menggunakan *confusion matrix*. Hasil penelitian ini diharapkan dapat bermanfaat untuk membantu melakukan riset atas opini masyarakat pada Twitter mengenai Pilpres 2019 yang mengandung sentimen positif, negatif atau netral.

METODE PENELITIAN

Metode penelitian yang digunakan pada penelitian ini terdiri atas beberapa tahap proses. Tahap awal merupakan akuisisi data yang diperoleh dari pengambilan data berupa *tweet* dari Twitter. Tahap *pre-processing*, yaitu terdiri dari *case folding*, normalisasi data *tweet*, *filtering*, mengubah kata baku, *stopword* serta *stemming*. Setelah dilakukkan *stemming*, tahap selanjutnya merupakan *processing*. *Processing* dilakukan dengan klasifikasi *Lexicon Based*, yaitu klasifikasikan data kedalam bentuk positif, negatif atau netral. Kemudian, *processing* dengan klasifikasi *Naïve Bayes* yang terdiri dari tahap

pelatihan dan pengujian data. Setelah itu, dilakukan evaluasi data dari hasil klasifikasi tersebut. Tahap akhir dari proses ini adalah visualisasi hasil data menggunakan *package shiny* pada RStudio. Akuisisi data (*tweets*) menggunakan API Search Twitter. Pada bahasa pemrograman R tersedia *package* untuk mengambil *tweets*. Tahap akuisisi terdiri dari proses pembuatan koneksi dengan API Search Twitter dan proses pengambilan datanya. Untuk mengambil data penelitian dalam penelitian ini *tweets* sesi R harus dihubungkan dengan API *Search* Twitter.

Preprocessing Data Twitter

Preprocessing data bertujuan untuk mentransformasi data ke dalam suatu format agar bisa lebih mudah dipahami. Tahap ini merupakan tahap penting dalam analisis sentimen. Pada tahap ini, tweet mengalami kapitalisasi dan pembersihan dari komponen @, RT, url, serta komponen lainnya. Data yang telah diambil dari Twitter tentunya perlu diekstraksi guna menghilangkan variabel yang tidak berguna. Merujuk pada penelitian sebelumnya yang dilakukan oleh [8] pada proses ini dilakukan *case folding*, normalisasi data *tweet*, *filtering*, mengubah kata baku, *stopword* dan *stemming* dengan penjelasan sebagai berikut :

1) *Case folding* atau kapitalisasi merupakan tahap pengubahan huruf pada data tweets menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap sebagai delimiter. Perubahan dilakukan dengan memeriksa ukuran setiap karakter dari awal sampai akhir karakter. Jika ditemukan karakter yang menggunakan huruf kapital, maka huruf tersebut akan diubah menjadi huruf kecil

2) Normalisasi Data Tweet. Data tweets yang didapat dari Twitter seringkali mengandung komponen yang tidak diperlukan seperti delimiter sehingga perlunya penghapusan komponen. Dalam R, penghapusan atau pernormalisasian dapat menggunakan fungsi pada *package tokenizers* dan *package textclean*. Proses pernormalisasian dilakukan dengan menghapus komponen emoji dan html pada data tweets, komponen ini tidak dibutuhkan dalam proses sentimen. Sintaks yang digunakan dalam R adalah :

```
tweets <- tweets %>%
  replace_emoji(.) %>%
  replace_html(.)
```

3) *Filtering*. *Filtering* adalah tahap menghilangkan tweets yang duplikat atau ganda menjadi satu. Terkadang, data yang diambil pada Twitter dapat berupa data duplikat atau ganda karena adanya *retweet* atau *repost tweet*. Maka dari itu dengan tahapan *filtering* ini, dapat menjadikan data tweets tidak ada yang ganda. Pada penelitian ini awalnya data yang di dapat dari tahapan akuisisi data adalah sebanyak 5818 tweets, namun setelah di *filtering* data menjadi sebanyak 1090 tweets.

4) Mengubah kata baku. Mengubah kata-kata pada tweets yang tidak baku atau kata-kata cakapan (*slang*) menjadi kata-kata yang baku sesuai dengan Kamus Besar Bahasa Indonesia. Tujuan proses ini sama dengan proses *stemming*.

5) *Stopword.* Pada tahap ini, kumpulan data tweets yang telah melewati tahap ubah kata baku akan melalui tahap menghapus kata-kata yang tidak perlu (*stopword*). Setiap kata pada data tweets akan diperiksa. Jika terdapat kata sambung, kata depan, kata ganti atau kata yang tidak ada hubungannya dalam analisis sentimen, maka kata tersebut akan dihilang-kan. Langkah-langkah membuang kata-kata yang tidak perlu adalah sebagai berikut: a) Hasil proses ubah kata baku akan disbanding-an dengan daftar *database* katakata yang tidak perlu b) Dilakukan pengecekan apakah terdapat kata sama dengan daftar atau tidak. c)Jika terdapat kata yang sama pada daftar *database*, maka kata yang sama tersebut akan dihilangkan.

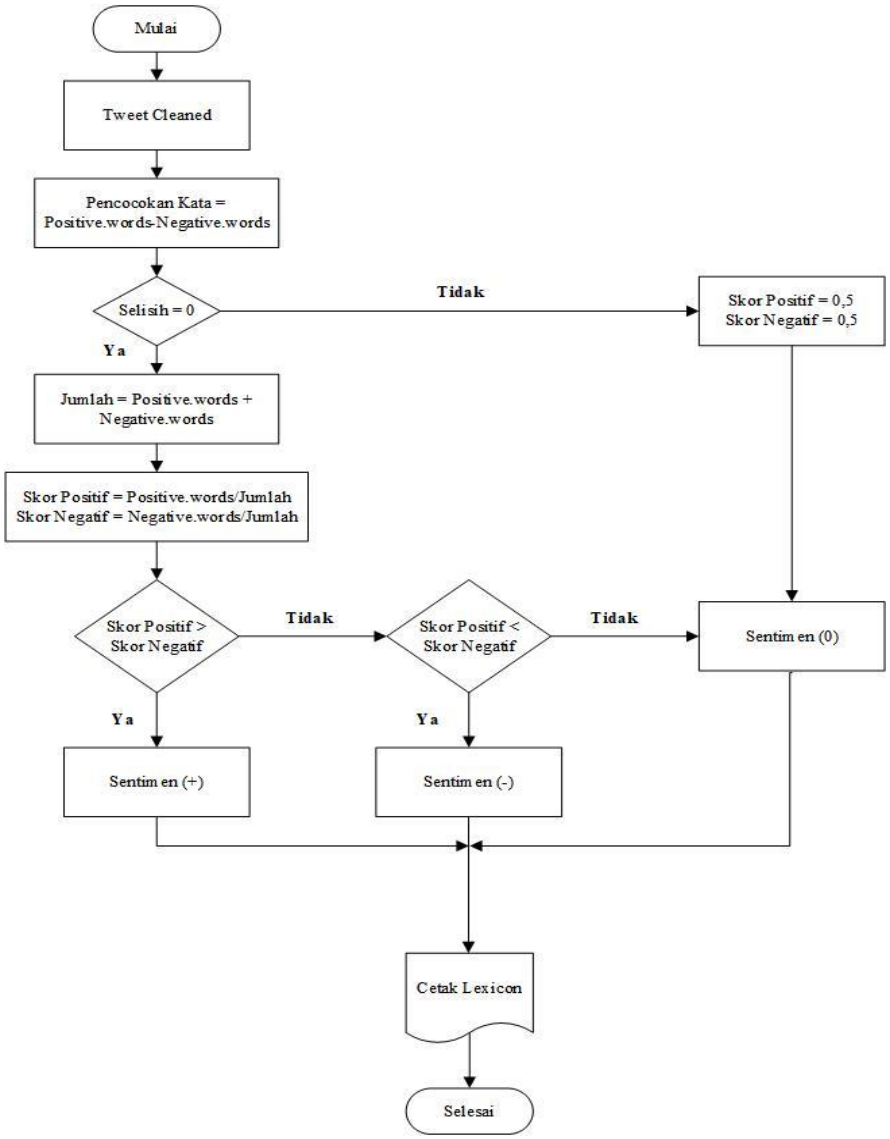
6) *Stemming*. *Stemming* adalah proses mengubah kata menjadi bentuk kata dasarnya dengan cara menghilangkan imbuhan-imbuhan pada kata dalam suatu dokumen. Algoritma *stemming* yang digunakan dalam penelitian ini adalah Algoritma Nazief dan Adriani.

Klasifikasi Data dengan Lexicon Based

Klasifikasi sentimen dengan *Lexical Based* adalah klasifikasi berdasarkan kata positif, kata negatif ataupun netral yang ada pada tweets yang telah dibersihkan.

Klasifikasi ini telah dicocokkan dengan kata-kata yang terdapat dalam kamus *Lexicon* Bahasa Indonesia. Jika tweets memiliki kata positif, maka akan digolongkan pada sentimen tweet positif. Jika tweets memiliki kata negatif, maka akan digolongkan pada sentimen tweet negatif. Namun pada kasus lain jika kedua kata ini bernilai sama, maka digolongkan dalam tweet netral. Bagan alur dari klasifikasi ini dapat dilihat pada Gambar 1. Klasifikasi menggunakan metode *Lexicon Based* dari data tweets sebanyak 1091 menghasilkan sentimen positif sebanyak 484, sentimen negatif sebanyak 392 dan sentimen netral sebanyak 214.

Berikut akan ditampilkan proses klasifikasi kalimat menggunakan metode *Lexicon Based*. Jumlah data yang ditampilkan sebanyak 17 data, baris disebelah kanan merupakan hasil sentimen yang dibagi kedalam sentimen positif, negatif atau netral berdasarkan kamus *Lexicon* yang tersedia.

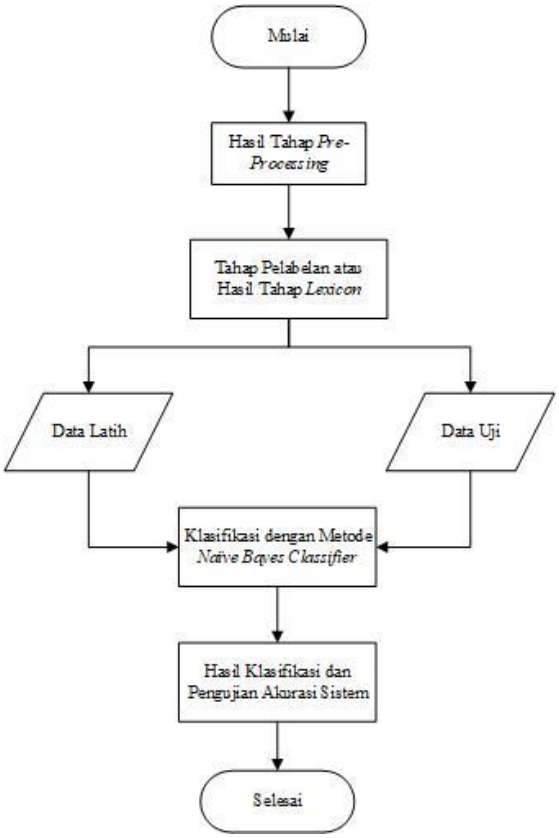


Gambar 1. Alur *Lexicon Based*

Klasifikasi Data dengan *Naive Bayes Classifier*

Tahap Selanjutnya adalah tahap klasifikasi, yaitu tahap pengklasifikasian untuk menentukan apakah data yang akan diuji termasuk kedalam sentimen positif atau negatif. Pada tahap ini digunakan sebuah metode yaitu *Naïve Bayes Classifier* dengan menggunakan pelabelan terlebih dahulu. Metode ini terdiri dari 2 proses, yaitu proses pelatihan dan proses pengujian.

Gambar 2 merupakan bagan alur dari klasifikasi. Adapun tahapan klasifikasi *Naïve Bayes Classifier* dengan menggunakan pelabelan.



Gambar 2. Alur *Naive Bayes Classifier*

Tabel 1. Kriteria Label Teks Tweets

Label	Kriteria
Positif	Berisi kata-kata "menang", "jujur", "dukung", "selamat"
Negatif	Berisi kata-kata "curang", "tolak", "bodoh"

Proses Pelabelan pada tahapan klasifikasi *Naive Bayes Classifier* adalah sebagai berikut:

1. Penentuan Kategori Sentimen Teks Tweets

Pada tahap ini teks tweets dibaca satu per satu kemudian ditentukan kriteria kategorinya berdasarkan kata-kata tertentu (kata kunci) yang muncul pada teks tweets. Kriteria label pada penelitian ini ditunjukkan oleh Tabel 1.

2. Labelisasi Tweets

Setelah menganalisis, penelitian dilanjutkan ke tahap labelisasi. Pada tahap labelisasi tweets diberikan 1 dari 2 label ("positif" atau "negatif"). Labelisasi dilakukan secara manual dengan cara menuliskan labelnya pada kolom tambahan yang dibuat pada berkas CSV hasil dari data *pre-processing* yang sudah di *stemming*.

Kedua label yang digunakan dalam penelitian ini dijelaskan sebagai berikut: (1) Label positif, dilabelkan pada teks tweets dengan

kriteria berupa dukungan maupun pembelaan. (2) Label negatif, dilabelkan pada teks tweets yang mengandung kata-kata ejekan, sindiran maupun hinaan.

3. Hasil Labelisasi Tweets

Adapun hasil pelabelan dai 1090 data tweets menghasilkan 807 teks tweets terlabeli positif dan 283 teks tweets terklasifikasi negatif. Beberapa contoh teks tweets beserta labelnya dituangkan ke dalam Tabel 2.

4. Transformasi

Setelah selesai pelabelan teks tweets, teks tersebut harus diubah menjadi bentuk yang dapat digunakan oleh komputer (atau lebih tepatnya *machine learning* dalam hal ini). Pengubahan bentuk tersebut dikenal sebagai vektorisasi. Dalam lingkup *text mining*, vektorisasi adalah proses membuat vektor dengan nilainya berupa angka-angka kemunculan term (yaitu kata unik) dalam dokumen [9]. Untuk melakukan vektorisasi maka teks tweets harus dipisah menjadi kata per kata. Pemisahan teks tweets menjadi kata-kata disebut dengan tokenisasi. Tokenisasi adalah proses pemecahan dokumen menjadi komponen-komponen individual [10]. Dalam hal ini komponen individual tersebut yaitu token adalah kata dari tiap teks tweets. Tujuan tokenisasi adalah agar dapat digunakan oleh model *machine learning* (Hari, 2015). Dalam penelitian ini token-token dibuat dalam bentuk *document term matrix*. Untuk membuat *document term matrix* digunakan perintah berikut:

```
dtm_tweets<-
DocumentTermMatrix(VCorpus(VectorSo
urce(tweets$Teks)))
```

Keterangan :

a. DocumentTermMatrix() adalah fungsi untuk membuat document term matrix dari sebuah corpora, yang disediakan oleh *package tm.*

b. VCorpus (VectorSource(tweets$teks)) adalah parameter berupa fungsi untuk membuat corpora volatil. Pada penelitian ini hanya variabel teks yang dijadikan corpora, sehingga dituliskan teks tweets$teks sebagai vektor sumber.

c. VectorSource() adalah fungsi untuk membuat sebuat vektor sumber.

Tabel 2. Contoh Teks Tweets yang Sudah Dilabeli

Label	Teks Tweets
Positif	selamat menang pilpres moga pimpin amanah rakyat indonesia
Negatif	tidak gratis bohong curang pilpres bentar lagi laknat dunia turun

5. Implementasi Algoritma Pengklasifikasi *Naïve Bayes*

Setelah selesai melakukan tahap transformasi, analisis dilanjutkan ke tahap Implementasi Algoritma Pengklasifikasi *Naïve Bayes*. Pada tahap *processing* diimplementasikan algoritma pengklasifikasi *Naïve Bayes*.

Adapun tahap *processing* pada penelitian ini terdiri dari 3 proses, yaitu **1)** Membuat Set Data Latih dan Uji. Untuk membuat set data latih maupun set data uji, objek dtm_tweets terlebih dahulu "dibagi" menjadi 2 buah objek menggunakan 2 perintah berikut:

```
dtm_tweets_latih <- dtm_tweets[1:845, ]
dtm_tweets_uji <- dtm_tweets[846:1090, ]
```

Keterangan:
[1:845,] dan [846:1090,] menunjukkan nomor baris teks tweets dalam objek dtm_tweets. Dengan menulis 2 perintah di atas, terbentuklah 2 buah objek matriks berisi data latih sebanyak 845 dan data uji sebanyak 245.

2) Melakukan Latih pada Set Data Latih

Setelah set data latih dan set data uji dibuat, maka dilanjutkan ke proses latih. Pada proses latih diimplementasikan algoritma pengklasifikasi *Naïve Bayes*. Algoritma pengklasifikasi *Naïve Bayes* akan menggunakan kemunculan tiap *term* (kata unik) dalam teks tweets untuk menghitung pembuatan set data latih dan set data uji, melakukan data latih (membangun model pengklasifikasi menggunakan algoritma *Naïve Bayes*), dan diakhiri dengan klasifikasi (mengimplementasikan model pengklasifikasi pada set data uji teks tweets). Berikut adalah proses-proses tahapan implementasi algoritma pengklasifikasi *Naïve Bayes*:

probabilitas klasifikasi teks tweets. Dengan kata lain, proses latih akan menghasilkan model pengklasifikasi teks tweets untuk penelitian ini.

Proses latih pada algoritma pengklasifikasi *Naïve Bayes* terdiri dari 2 buah perhitungan, yaitu perhitungan untuk mencari probabilitas kemunculan label (kategori sentimen) dan perhitungan untuk mencari probabilitas kemunculan tiap *term* untuk tiap klasifikasi. Proses latih dapat dituliskan dengan 2 persamaan berikut:

$$P(\text{Label}_i) = \frac{\text{Jumlah Label}_i}{\text{Panjang Data Latih}} \quad (1)$$

$$P(\text{Term}_n|\text{Label}_i) = \frac{\text{Frekuensi Kemunculan Term}_n \text{ pada Teks dengan Label}_i}{\text{Jumlah Label}_i} \quad 2)$$

3) Melakukan Klasifikasi

Setelah model pengklasifikasi dibuat, maka proses pengklasifikasian dapat dilakukan. Berdasarkan teorema *Naïve Bayes*, probabilitas kategori teks tweets dituliskan pada persamaan 3.

$$P(V_{MAP}) = \frac{P(term_1, term_2, term_3, \ldots, term_n | label_i) P(label_i)}{P(term_1, term_2, term_3, \ldots, term_n)} \quad (3)$$

Nilai $P(term_1, term_2, term_3, \ldots, term_n)$ konstan untuk semua $label_i$, sehingga persamaan di atas dapat dapat disederhanakan menjadi persamaan 4.

$$V_{MAP} = \prod_{x=1}^{n} P(term_x | label_i) \frac{1}{Z} P(label_i) \quad (4)$$

Keterangan:

a. $\prod P(term_x | label_i)$ n x=1 adalah hasil perkalian sekuensial dari nilai probabilitas *term* pertama yang muncul pada label pertama hingga nilai probabilitas *term* ke-n yang muncul pada label terakhir.

b. Z adalah faktor penyekala yang mengubah probabilitas menjadi persamaan di atas. Nilai Z dihitung secara otomatis oleh algoritma.

HASIL DAN PEMBAHASAN

Visualisasi data dalam bentuk histogram hasil analisis sentimen menggunakan lexicon based dan naïve bayes classifier dapat dilihat pada Gambar 3. Gambar visualisasi data dalam bentuk *wordcloud* dapat dilihat pada Gambar 4.

Gambar 3. Tampilan Halaman Visualisasi Histogram

Gambar 4. Tampilan Halaman Visualisasi *Wordcloud*

Pengujian Sistem dengan *Confusion Matrix*

Pengujian akurasi klasifikasi dilakukan untuk mengetahui tingkat akurasi klasifikasi data *tweets* yang dilakukan secara manual menggunakan Labelisasi dan dengan metode *Lexicon Based* dengan klasifikasi data *tweets* yang dilakukan oleh sistem dengan menggunakan metode *Naïve Bayes Classifier*. Pengujian dilakukan dengan menggunakan *confusion matrix* yaitu sebuah matrik dari prediksi yang akan dibandingkan dengan kelas yang asli dari data inputan. Pengujian dilakukan menggunakan 245 data uji yang sudah diberi label. Data uji yang telah diklasifikasikan secara manual akan dibandingkan dengan hasil klasifikasi yang dilakukan oleh sistem menggunakan metode *Naïve Bayes Classifier*. Hasil pengujian akurasi klasifikasi data *tweets* dapat dilihat pada Tabel 3 dan 4.

Akurasi = (TNegatif + TNetral + TPositif) / Data Uji

Akurasi = (78 + 26 + 54) / 245

Akurasi = 158 / 245 = 0,6449 * 100% = 64,49%

Data pengujian akurasi yang digunakan pada Tabel 3 dari 845 data latih dan 245 data uji, sistem berhasil mengklasifikasi 158 data dengan tepat dan 87 data dengan keliru. Berdasarkan pengujian akurasi, didapatkan hasil akurasi klasifikasi data *tweets* dari sistem analisis sentimen dengan menggunakan metode *Lexicon Based* dan diuji dengan sistem *Naïve Bayes Classifier* sebesar 64,49%.

Tabel 3. Tabel *Confusion Matrix Lexicon* dengan Sistem

Prediksi Sistem	***Lexicon***		
	Negatif	**Netral**	**Positif**
Negatif	78	21	8
Netral	20	26	18
Positif	11	9	54

Tabel 4. Tabel *Confusion Matrix* Labelisasi dengan Sistem

Jumlah Data Uji: 245	**Sentimen Hasil Analisis Negatif**	**Sentimen Hasil Analisis Positif**
Sentimen Asli Negatif	(TrueNegatif) 25	(FalsePositif) 14
Sentimen Asli Positif	(FalseNegatif) 19	(TruePositif) 187

1. Akurasi

Akurasi = (TNegatif +TPositif) / Data Uji

Akurasi = (25+187) / 245

Akurasi = 212 / 245 = 0,8653 * 100% = 86,53%

2. Kesalahan Sistem

Kesalahan Sistem = 1 – Akurasi

Kesalahan Sistem = 1 – 0,8653 = 0,1347 * 100% = 13,47%

3. Presisi

Presisi Positif = (TP) / (TP + FP)

Presisi Positif = 187 / (187+14)

Presisi Positif = 187 / 201 = 0,9303 * 100% = 93,03%

Presisi Negatif = (TN) / (TN + FN)

Presisi Negatif = 25 / (25+19)

Presisi Negatif = 25 / 44 = 0.5681 * 100% = 56,81%

4. Recall/Sensitifity

Recall = (TP) / (TP+FN)

Recall = 187 / (187+19)

Recall = 187 / 206 = 0.9077 * 100% = 90,77%

5. Specificity

Specificity = (TN) / (TN+FP)

Specificity = 25 / (25+14)

Specificity = 25 / 39 = 0,6410 * 100% = 64,10%

Data pengujian akurasi yang digunakan pada Tabel 4 dari 845 data latih dan 245 data uji, sistem berhasil mengklasifikasi 212 data dengan tepat dan 33 data dengan keliru. Berdasarkan pengujian akurasi, didapatkan hasil akurasi klasifikasi data *tweets* dari sistem analisis sentimen dengan menggunakan Labelisasi dan diuji dengan sistem *Naïve Bayes Classifier* sebesar 86,65% dengan kesalahan sistem sebesar 13,47%, presisi positif sebesar 93,03% serta presisi negatif sebesar 56,81%, *recall* sebesar 90,77% dan *specificity* sebesar 64,10%. Kesimpulan yang diperoleh dari pengujian akurasi menggunakan *confusion matrix* adalah bahwa analisis sentimen menggunakan Labelisasi dan *Naïve Bayes Classifier* dapat digunakan sebagai metode pengklasifikasian pada analisis sentimen karena tingkat akurasinya yang lebih besar dibandingkan dengan menggunakan metode *Lexicon Based*.

KESIMPULAN DAN SARAN

Hasil klasifikasi menggunakan metode *Lexicon Based* menghasilkan klasifikasi positif sejumlah 484, negatif 392 dan netral 214 dari data *tweets* sebanyak 1090. Pada metode ini klasifikasi terbanyak adalah pada sentimen positif. Klasifikasi menggunakan metode *Naïve Bayes Classifier* dan pelabelan manual dengan data latih sebanyak 845 dan data uji sebanyak 245, menghasilkan klasifikasi sentimen positif sebanyak 206 dan negatif sebanyak 39. Pada metode ini hasil klasifikasi terbanyak adalah pada sentiment positif. Pada penelitian analisis sentimen ini didapatkan hasil akurasi yang diuji oleh *Confusion Matrix*, klasifikasi data *tweets* dari sistem analisis sentimen dengan menggunakan metode *Lexicon Based* sebesar

64,49% pada data *testing* dan pada data *training* sebanyak 94,2% serta dengan menggunakan Labelisasi dan *Naive Bayes Classifier* sebesar 86,53% pada data *testing* dan data *training* sebesar 94,08%. Maka berdasarkan hasil analisis sentimen pada penelitian ini dapat disimpulkan bahwa opini masyarakat melalui Twiter mengenai Pilpres 2019 menggunakan 2 metode yaitu *Lexicon Based* dan *Naïve Bayes Classifier*, hasil yang mendapatkan tingkat akurasi paling tinggi yaitu dengan Labelisasi *Naïve Bayes Classifier* dan dari kedua metode tersebut hasil dari klasifikasi sentimen memiliki kecenderungan yang sama yaitu sentimen positif.

Pengembangan lebih lanjut dapat dilakukan seperti data yang ditarik dari Twitter dalam jumlah yang besar, menggunakan algoritma yang berbeda, selain melakukan *classification* juga melakukan *clustering* untuk pemisahan data dalam jumlah besar, serta adanya fitur *real-time* pada visualisasi data dan dapat diaskes kapan saja. Pada pengembangan lebih lanjut dari sistem yang sudah dibuat, diharapkan aplikasi ini dapat dikembangkan dengan menggunakan metode lain yang lebih baik dan dengan mengikuti perkembangan teknologi informasi.

DAFTAR PUSTAKA

[1] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi, dan H. E. Williams, "Stemming Indonesian: A confix-stripping approach", *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 6, no. 4, hal. 1 – 33, 2007.

[2] X. Ding, B. Liu, dan P. S. Yu, "A holistic lexicon-based approach to opinion mining", Dalam Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008.

[3] G. A. Buntoro, T. B. Adji, dan A. E. Permanasari ",Sentiment analysis Twitter dengan kombinasi *lexicon based* dan *double propagation*," dalam *CITEE 2014*, 2014, hal. 39 – 43.

[4] Falahah dan D. W. A. Nur, "Pengembangan aplikasi *sentiment analysis* menggunakan metode Naive Bayes (Studi kasus sentiment analysis dari media Twitter)", Dalam Seminar Nasional Sistem Informasi Indonesia, 2015 Hal. 335 – 340.

[5] I. Kusumawati, "Analisa sentimen menggunakan *lexicon based* kenaikan harga rokok pada media sosial Twitter,",Skripsi Sarjana, Universitas Muhammadiyah Surakarta, Surakarta, 2017.

[6] G. A. Buntoro, "Analisis sentimen calon Gubernur DKI Jakarta 2017 di Twitter", *INTEGER: Journal of Information Technology*, vol. 2, no. 1, hal. 32 – 41, 2017.

[7] M. M. Munir, M. A. Fauzi, dan R. S. Perdana, "Implementasi metode

backpropagation neural network berbasis lexicon based features dan bag of words untuk identifikasi ujaran kebencian pada Twitter," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 10, hal. 3182 – 3191, 2018.

[8] S. K. Ravindran dan V. Garg, Vikram, *Mastering Social Media Mining with R.* Packt Publishing Ltd. UK, 2015.

[9] W. Baugh, 2013. *https://stackoverflow.com/a/17054702*.

[10] B. Lantz, *Machine Learning With R.* Birmingham: Packt Publishing Ltd, 2015.