

**ANALISIS SENTIMEN DATA TWITTER MENGGUNAKAN NAIVE BAYES DENGAN NEGATION HANDLING**

**DEPARTEMEN ILMU KOMPUTER**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**INSTITUT PERTANIAN BOGOR**

**BOGOR**

**2016**

**ASWINDA PRIMA PUTRA**

**PERNYATAAN MENGENAI SKRIPSI DAN**

**SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA**

Dengan ini saya menyatakan bahwa skripsi berjudul Analisis Sentimen Data Twitter menggunakan Naive Bayes dengan Negation Handling pada Data Twitter Bahasa Indonesia adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, November 2016

*Aswinda Prima Putra*

NIM G64144007

**ABSTRAK**

ASWINDA PRIMA PUTRA. Analisis Sentimen Data Twitter menggunakan Naive Bayes dengan Negation Handling. Dibimbing oleh JULIO ADISANTOSO.

Penelitian ini melakukan klasifikasi sentimen dalam 3 jenis yaitu positif, negatif dan netral menggunakan metode klasifikasi Multinomial Naïve Bayes. Analisis sentimen dalam penelitian ini menggunakan data twitter berbahasa Indonesia dengan data seimbang. Sentimen dalam data latih dilakukan secara subyektif maka *clustering* dengan menggunakan *k-means* akan digunakan untuk melihat bagaimana data mengelompok untuk dilakukan pemberian sentimen ulang pada data tweet yang berbeda dengan label *cluster*. Setelah dilakukan pemberian sentimen ulang, data yang masih berbeda dengan cluster akan dihapus dan didapatkan peningkatan akurasi sebesar 20.29%. Negation Handling memiliki peran untuk menambah akurasi dari klasifikasi. Dalam penelitian ini Negation Handling menunjukkan peningkatkan akurasi sebesar 3.45%. Dalam penelitian ini juga menambahkan fitur untuk menangani kata yang hilang akibat bentuk negasi “tidak” + kata, dan mendapatkan peningkatkan sebesar 3.59% akurasi dalam menggunakan Negation Handling dengan penambahan fitur kata yang hilang.

Kata kunci**:** analisis sentimen, klasifikasi, naïve bayes, negation handling, twitter

**ABSTRACT**

*ASWINDA PRIMA PUTRA. Sentiment Analysis Twitter Data with Rocchio Classification for Indonesian Language. Supervised by JULIO ADISANTOSO.*

*This research is trying to classify the sentiment into three types: positive, negative and neutral using Multinomial Naïve Bayes. This research will be using Twitter data in Bahasa Indonesia with balanced sentiment data. Sentiment in the data train is done subjectively then clustering with k-means will be used to see how the data is grouped for re-labeling sentiment tweet which have different sentiment with cluster label. After re-labeling sentiment, the data which have different sentiment with cluster label will be removed and obtained an increase in accuracy of 20.29%. Negation Handling has a role in increasing the accuracy of classification. In this research Negation Handling show increasing accuracy of 3.45%. In this research also add feature to handle missing word due to bootstrap word with “tidak” + word, and show increasing accuracy of 3.59% with the using of Negation Handling with adding missing word feature.*

*Keywords: sentiment analysis, classification, Naïve Bayes, Negation Handling, Twitter.*

Skripsi  
sebagai salah satu syarat untuk memperoleh gelar  
Sarjana Komputer  
pada  
Departemen Ilmu Komputer

**ASWINDA PRIMA PUTRA**

**ANALISIS SENTIMEN DATA TWITTER MENGGUNAKAN NAIVE BAYES DENGAN NEGATION HANDLING**

**DEPARTEMEN ILMU KOMPUTER**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**INSTITUT PERTANIAN BOGOR**

**BOGOR**

**2016**

Penguji : 1

2

Judul Skripsi : Analisis Sentimen Data Twitter menggunakan Naive Bayes dengan Negation Handling

Nama : Aswinda Prima Putra

NIM : G64144007

Disetujui oleh

|  |
| --- |
| Ir Julio Adisantoso, MKom  Pembimbing |

Diketahui oleh

Dr Ir Agus Buono, MSi MKom

Ketua Departemen

Tanggal Lulus:

**PRAKATA**

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa ta’ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Agustus 2016 ini ialah Analisis Sentimen Data Twitter menggunakan Naive Bayes dengan Negation Handling pada Data Twitter Bahasa Indonesia.

Terima kasih penulis ucapkan kepada Bapak Ir. Julio Adisantoso MKom selaku pembimbing. Ungkapan terima kasih juga disampaikan kepada kedua orang tua saya. Terimakasih juga disampaikan kepada teman satu bimbingan serta teman-teman ekstensi Ilkom IPB angkatan 9 yang telah membantu dalam menyelesaikan penelitian ini.

Semoga karya ilmiah ini bermanfaat.

Bogor, Desember 2016

*Aswinda Prima Putra*

**DAFTAR ISI**

[PENDAHULUAN 1](#_Toc472518625)

[Latar Belakang 1](#_Toc472518626)

[Tujuan Penelitian 2](#_Toc472518627)

[Manfaat Penelitian 2](#_Toc472518628)

[Ruang Lingkup Penelitian 2](#_Toc472518629)

[METODE 2](#_Toc472518630)

[Praproses 3](#_Toc472518631)

[Indexing 4](#_Toc472518632)

[Tokenisasi 5](#_Toc472518633)

[Normalisasi Kata 5](#_Toc472518634)

[Negation Handling 5](#_Toc472518635)

[Penghapusan Stopwords 6](#_Toc472518636)

[Pemilihan Fitur 7](#_Toc472518637)

[Klasifikasi 7](#_Toc472518638)

[Lingkungan Pengembangan 9](#_Toc472518639)

[HASIL DAN PEMBAHASAN 10](#_Toc472518640)

[Praproses 10](#_Toc472518641)

[Indexing 11](#_Toc472518642)

[Negation Handling 11](#_Toc472518643)

[Pembuangan Stopword 12](#_Toc472518644)

[Seleksi fitur 12](#_Toc472518645)

[Evaluasi 13](#_Toc472518646)

[Clustering 15](#_Toc472518647)

[Evaluasi 16](#_Toc472518648)

[KESIMPULAN DAN SARAN 19](#_Toc472518649)

[Kesimpulan 19](#_Toc472518650)

[Saran 19](#_Toc472518651)

[DAFTAR PUSTAKA 20](#_Toc472518652)

# 

# PENDAHULUAN

## Latar Belakang

*Microblogging* saat ini telah menjadi alat komunikasi yang sangat populer di kalangan pengguna internet. Jutaan penggunanya membagikan opini tentang bebagai macam aspek dari kehidupan ataupun membahas isu-isu saat ini. Oleh karena itu *microblogging* merupakan sumber data yang sangat kaya untuk melakukan *opinion mining* dan analisis sentimen (Pak dan Paroubek. 2010). Berbagai macam perusahaan sering menggunakan *survey online* atau dengan kertas untuk mengumpulkan opini dari penggunanya. Namun dengan kemunculan media sosial, orang cenderung lebih memilih mengutarakan pendapatnya melalui facebook, twitter atau media sosial yang lainnya (Anwar *et al.* 2015).

Salah satu situs *microblogging* yang populer adalah Twitter. Twitter merupakan situs web yang dioperasikan oleh Twitter, Inc yang memberikan jaringan sosial berupa microblog serta memiliki karakteristik dan format yang unik dengan simbol ataupun aturan khusus. Disebut *microblog* karena penggunanya hanya dapat mengirim dan membaca pesan *blog* seperti pada umumnya dengan batas maksimal sejumlah 140 karakter, pesan tersebut dikenal dengan tweet (Zhang et al. 2011). Twitter digunakan dalam penelitian ini karena sifat dari semua tweetnya bersifat publik. Berbeda dengan sosial media lain seperti facebook yang karakteristik dari *posting* nya dapat dirubah menjadi privat atau hanya di lingkaran teman nya saja.

Sentimen analisis atau opinion mining adalah studi komputasional dari opini-opini orang, sentimen dan emosi melalui entitas dan atribut yang dimiliki yang diekspresikan dalam bentuk teks Liu (2012). 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 et al. 2002). Salah satu teknik pembelajaran mesin untuk analisis sentimen adalah *Naïve Bayes* Classifier (NBC). NBC merupakan teknik pembelajaran mesin yang berbasis probabilistik. NBC adalah metode sederhana tetapi memiliki akurasi serta performansi yang tinggi dalam pengklasifikasian teks (Routray et al. 2013). Xhemali et al. (2009) melakukan perbandingan antara tiga metode, metode-metode tersebut adalah *Naïve Bayes*, Pohon Keputusan, dan Neural Networks. Hasil penelitian secara keseluruhan menunjukkan bahwa *Naïve Bayes* classifier adalah pilihan terbaik untuk pelatihan domain. Penggunaan model *Naïve Bayes Classifier* ini dikarenakan proses yang sederhana dan mudah diaplikasikan pada berbagai keadaan sehingga tidak akan mengalami kegagalan secara keseluruhan pada hasilnya (Manning et al. 2008).

Penelitian Pang et al (2002) menunjukkan klasifikasi menggunakan naïve bayes menghasilkan akurasi yang kurang baik dibandingkan dengan menggunakan klasifikasi seperti SVM dan maximum entropy. Dalam penelitiannya klasifikasi menggunakan SVM menghasilkan hasil paling baik diantara Maximum entropy dan Naïve Bayes. Dalam penelitian Adityawan (2014) menggunakan data Twitter berbahasa Indonesia juga masih menghasilkan akurasi yang kurang baik yaitu sebesar 71.06% Multinomial Naïve Bayes dan 66.42% untuk Bernoulli, sedangkan dalam penelitian Narayanan (2013) klasifikasi Naïve Bayes dengan penambahan fitur Negation Handling, seleksi fitur Mutual Information dan N-Gram menunjukkan akurasi yang lebih baik daripada menggunakan klasifikasi SVM. Dalam penelitian tersebut penambahan fitur Negation Handling memberikan peningkatan akurasi paling banyak dibandingkan dengan seleksi fitur Mutual Information dan N-Gram sebesar 10%. Maka dari itu klasifikasi Naïve bayes dengan penambahan fitur Negation Handling akan diterapkan pada penelitian ini dengan menggunakan data Twitter berbahasa Indonesia.

**Perumusan Masalah**

Perumusan masalah pada penelitian ini adalah:

1. Bagaimana mengimplementasikan Metode Negation Handling pada analisis sentimen Twitter berbahasa Indonesia ?
2. Apakah metode Negation Handling dapat meningkatkan akurasi analisis sentimen dibandingkan tanpa menggunakan Negation Handling ?

## Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Mengimplementasikan analisis sentimen dengan menggunakan metode Negation Handling pada data Twitter berbahasa Indonesia.
2. Membandingkan akurasi dari analisis sentimen dengan Negation Handling dan tanpa menggunakan Negation Handling.

.

# Manfaat Penelitian

Hasil penelitian diharapkan dapat membantu pengguna yang ingin mengetahui isu tertentu dari data Twitter dan memberikan informasi apakah isu tersebut mengandung sentiment positif, negatif, atau netral.

## Ruang Lingkup Penelitian

Ruang lingkup penelitian ini menggunakan data Twitter berbahasa Indonesia dan tidak menangani tweet dengan polaritas yang ambigu.

# METODE

Penelitian ini terdiri atas 6 tahap yaitu praproses (pengumpulan data, pembagian data), indexing(tokenisasi, normalisasi kata, Negation Handling, pembuangan *stopword*), seleksi fitur, klasifikasi, evaluasi, dan clustering. Skema tahapan penelitian dapat dilihat pada Gambar 1.

Pengumpulan data Twitter

Pembagian data

1. Praproses

2. Indexing

Tokenisasi

Nomalisasi

Negation Handling

3. Seleksi Fitur

Data Latih

Data Uji

*Naïve Bayes*

4. Klasifikasi

5. Evaluasi Klasifikasi

6. *Clustering*

Akurasi < 75 %

ya

tidak

Pembuangan Stopword

Gambar Metode Penelitian

## Praproses

Data yang digunakan adalah data Twitter berbahasa Indonesia. Data yang digunakan diperoleh dari Twitter API melalui aplikasi yang telah dibuat oleh Hawksey (2013)*,* dengan memasukkan kata kunci untuk penyaringan data ke situs *tags.hawksey.info*. Data yang akan diambil adalah data *tweet* dengan kata kunci tertentu terkait isu yang akan diambil datanya. Kata kunci yang digunakan dalam penelitian ini adalah “IPB”, “Institut Pertanian Bogor”, “Bogor Agricultural University”, “Pilgub”, dan “Pilkada DKI”. Struktur data yang didapatkan dari Hawskey (2013) seperti pada Tabel 1.

Data yang akan digunakan dalam penelitian ini adalah data dengan atribut *text*. Data yang diperoleh masih berupa data yang belum ada sentimennya, untuk itu dilakukan pemberian sentimen secara manual. Pemberian sentimen dibagi menjadi 3 kelas, yaitu positif, negatif, dan netral. Selain data dari Hawskey (2013) data ditambahkan dengan data penelitian Adityawan (2014) sebanyak 1997 data dan Aziz (2013) sebanyak 27275 data.

Table Struktur Data Twitter

|  |  |
| --- | --- |
| Atribut | Keterangan |
| id\_str | id dari *post Twitter* |
| from\_user | *username* pemakai Twitter |
| text | *post* Twitter |
| created\_at | tanggal dan waktu *post* dibuat |
| geo\_coordinates | koordinat tempat user |
| Source | link profile user |
| profile\_image\_url | gambar profil dari user |
| user\_followers\_count | jumlah follower user |
| user\_friends\_count | jumlah teman user |
| user\_location | lokasi dari user |
| status\_url | link dari *post* Twitter |

Setelah semua data tweet terkumpul, kemudian dilakukan pembersihan data yaitu dengan cara menghilangkan tweet yang duplikat. Dalam proses ini *retweet* dengan simbol “RT” dianggap sebagai tweet duplikat jadi tweet tersebut akan dihapus. Tweet dengan Bahasa selain Bahasa Indonesia juga dihapus karena penelitian dikhususkan untuk data berbahasa Indonesia.

Pembagian data yang digunakan adalah data seimbang pada setiap sentimen karena pada penelitian Adityawan (2014) dengan data seimbang akurasi dari setiap sentimen akan seimbang atau tidak terlalu jauh. Data akan dibagi menjadi data latih dan data uji dengan presentasi 80% untuk data latih dan 20% untuk data uji pada setiap sentimennya. Pengambilan sampel data pada setiap sentimen dilakukan secara acak.

## Indexing

Indexing merupakan proses persiapan yang dilakukan terhadap dokumen sehingga dokumen siap untuk diproses. Proses indexing dibagi menjadi dua proses, yaitu document indexing dan term indexing. Dari term indexing akan dihasilkan koleksi kata yang akan digunakan untuk meningkatkan performansi pencarian pada tahap selanjutnya. Selain itu, teknik indexing ini juga dilakukan agar hasil yang diperoleh lebih baik. Karena kebanyakan tweet hanya berisi tautan dan tidak menunjukkan sentimen tertentu, dan penulisannya ditulis dalam bahasa asing yang bukan bahasa Inggris (Parikh dan Movassate, 2009). Tahap-tahap yang dilakukan didalam proses indexing meliputi tokenizing, pengahapusan stopwords, normalisasi kata, stemming, pembuatan term document matrix.

## Tokenisasi

Tokenisasi adalah proses untuk memotong dokumen menjadi bagian-bagian kecil yang disebtu token (Manning et al. 2009). Dalam proses tokenisasi karakter seperti tanda baca ataupun simbol juga akan dihapus. *Whitespace* (spasi, tab, *newline*) digunakan sebagai pemisah antar token yang akan dipotong.

Dalam penelitian ini proses tokenisasi akan dibagi menjadi dua bagian. Bagian pertama dalam proses tokenisasi adalah proses menghilangkan karakter diluar nilai ASCII 32 sampai 126 karena karakter diluar nilai ASCII tersebut bukan merupakan huruf latin dalam Bahasa Indonesia. Lalu proses selanjutnya adalah mengubah karakter menjadi huruf kecil dan menghapus URL. Dalam proses tokenisasi pertama ini tanda baca seperti “-)(\/=.,:;!?” tidak akan dihilangkan karena tanda baca tersebut akan digunakan dalam proses Negation Handling.

Proses tokenisasi yang kedua dilakukan setelah proses Negation Handling dan penghapusan *stopword*. Dalam proses ini akan dilakukan penyimpanan token unik ke dalam database “*bag-of-words*”. Dalam proses ini juga akan dilakukan penghitungan nilai tf positif, negatif, netral dan df dari setiap token unik.

## Normalisasi Kata

Normalisasi kata merupakan proses untuk mengubah kata tidak baku menjadi kata baku, karena dalam kata baku ambiguitas pelafalannya akan lebih kecil dibandingkan dengan kata tidak baku (Aziz. 2013). Tahap normalisasi kata juga dilakukan untuk mengurangi kata dengan arti yang sama, misal kata “ga”, “gak”, dan “nggak” mempunyai arti yang sama dengan “tidak”, jika tidak dilakukan normalisasi kata maka kata “ga”, “gak”, “nggak” dan “tidak” akan dianggap fitur yang berbeda.

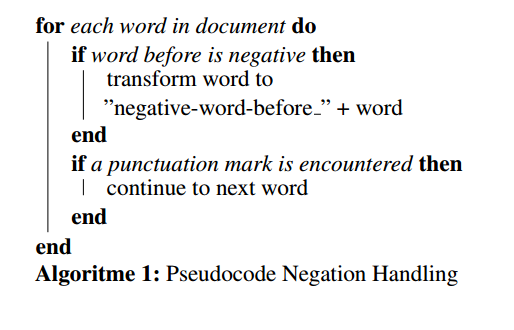
Pada penelitian ini dataset kata baku diperoleh dari penelitian Adityawan (2014) dengan jumlah data 3719 baris, Contoh daftar kata normalisasi dapat dilihat pada Lampiran 1.

## Negation Handling

Dalam penelitian Narayanan (2013), Negation Handling dapat meningkatkan akurasi dari klasifikasi secara signifikan yaitu sebesar 9.03%. Data yang digunakan dalam penelitian Narayanan (2013) dari dataset review film dari *Internet Movie Data base (IMDb)* dalam Bahasa Inggris dengan jumlah 25.000 data untu data laith dan 25.000 untuk data uji. Dalam penelitiannya sentiment yang digunakan yaitu sentimen positif dan negatif.

Setiap kata dala dokumen akan digunakan sebagai fitur, misalnya kata “baik” dalam kalimat “tidak baik” akan memberikan kontribusi sentiment positif daripada sentimen negatif karena kata “tidak baik” merupakan fitur yang tepisah sehingga memberikan kontribusi peluang sentimen masing-masing daripada memberikan satu kontribusi peluang yaitu negatif.

Untuk menyelesaikan masalah tersebut Narayanan (2013) menggunakan *state variable* atau *bootstrapping* untuk menangani negasi tersebut. Dalam penelitiannya *state variable* yang digunakan adalah “\_”. Jika ada kata “tidak” akan digabungkan dengan kata selanjutnya menjadi “tidak\_” + kata. Kata tersebut tidak akan digabungkan jika kata setelah “tidak” adalah tanda baca atau kata “tidak” lagi. Pseudocode yang digunakan dalam penelitian ini adalah seperti [1]



Dalam penelitan Narayanan (2013) juga menyebutkan kemungkinan jumlah fitur negasi tidak cukup banyak untuk melakukan klasifikasi yang benar, sehingga semisal ada kata “0baik” dalam dokumen positif, maka jumlah fitur “baik” dalam kelas positif akan bertambah dan juga fitur “tidak\_baik” dalam kelas negatif akan bertambah. Modifikasi tersebut dapat menambah akurasi sebesar 1% dalam penelitian Narayanan (2013).

Dalam penelitian ini proses Negation handling akan dilakukan setelah proses normalisasi karena kata seperti “ga”, “gak”, “g” yang kata bakunya “tidak” tidak akan diubah menjadi bentuk negasi jika kata tersebut belum dinormalisasi menjadi kata “tidak”.

## Penghapusan *Stopwords*

*Stopwords* adalah sebuah kata-kata dalam bahasa tertentu yang sangat umum dan memiliki nilai informasi nol (Feinerer et al. 2008). Penghapusan *stopword* tidak akan mengubah makna atau informasi dari dokumen. Beberapa contoh dari *stopword* dalam Bahasa Indonesia adalah yang, juga, dari, dia, kamu, aku, saya. Daftar kata *stopword* yang digunakan dalam penelitian ini didapatkan dari penelitian Tala (2003) sebanyak 759 kata.

Proses Negation Handling dilakukan sebelum memasuki tahap pembuangan *stopword*, karena dataset yang digunakan dari penelitian Tala (2003) memuat kata “tidak” sehingga jika proses Negation Handling dilakukan setelah proses pembuangan *stopword* maka kata “tidak” yang digunakan dalam Negation Handling akan dihapus.

## K-fold Cross Validation

Pembagian data uji dan data latih menggunakan metode k-fold cross validation. Menurut Meira dan Zaki (2014) k-fold cross validation adalah pembagian data dengan membagi dataset sebanyak k subset data menjadi data uji dan subset data lainnya menjadi data latih. Pada penelitian ini jumlah k yang digunakan adalah sebanyak 10. Kohavi (1995) telah melakukan penelitian terhadap beberapa pendekatan untuk meningkatkan akurasi, yaitu pendekatan Cross-validation( termasuk regular cross-validation, leave-one-out cross-validation, stratified cross-validation) dan bootstrap, dan diantara beberapa pendekatan tersebut 10-fold crossvalidation merupakan metode yang paling baik karena memberikan kontribusi yang baik untuk meningkatkan akurasi. Pada penelitian ini setiap subset data akan menggunakan data seimbang, sehingga jumlah sentimen positif, negatif dan netral setiap subset akan sama.

## Pemilihan Fitur

Pemilihan fitur merupakan proses pemilihan subset dari term yang mewakili informasi penting dari sebuah dokumen. Tujuan dari pemilihan fitur adalah mengurangi ukuran kosakata dan meningkatkan hasil dari klasifikasi (Manning et al. 2008). Pemilihan fitur dibagi menjadi dua yaitu *unsupervised feature selection* dan *supervised feature selection* (Garnes, 2009). *Supervised feature selection* yaitu metode yang menggunakan informasi kelas dalam data latihnya. Contoh dari *supervised feature selection* adalah MI dan *chi-square*. *Unsupervised feature selection* adalah sebuah metode pemilihan fitur yang tidak menggunakan informasi kelas dalam data latihnya ketika memilih fitur untuk klasifikasi. Salah satu contoh pemilihan fitur yang tidak menggunakan informasi kelas dalam pemilihan fiturnya adalah *Inverse document frequency* (IDF).

*Inverse document frequency* (IDF) adalah *inverse* dari nilai *Document frequency* (DF). Menurut Witten (1999) kata yang jarang atau paling sedikit muncul justru harus diperhatikan sebagai kata yang lebih penting dari pada kata yang paling sering muncul dalam dokumen. Nilai dari IDF dapat dihitung dengan formula

(1)

dengan N adalah jumlah seluruh dokumen dan adalah jumlah dokumen yang mengandung term tertentu. Pada penelitian ini pemilihan fitur IDF dipilih karena metode ini efisien, mudahdan memiliki hasil yang akurat (Robertson 2005).

## Klasifikasi

Klasifikasi merupakan proses untuk menentukan kelas dari suatu objek (Rachman 2011). Objek yang digunakan dalam penelitian ini adalah data tweet yang selanjutnya akan diklasifikasikan ke dalam tiga kelas yaitu kelas positif, negatif atau netral. Pada penelitian ini klasifikasi menggunakan metode berbasis peluang, yaitu *Multinomial Naïve Bayes.* Perhitungan peluang didapatkan dengan persamaan,

dengan parameter 𝑃(𝑐) merupakan peluang kelas c, 𝑃(𝑡𝑘 |𝑐) adalah peluang token 𝑡𝑘 muncul pada dokumen c, dan 𝑛𝑑 adalah jumlah token unik pada dokumen (Manning et al. 2008). Kemudian nilai P(c) dan 𝑃(𝑡𝑘 |𝑐) didapatkan dari persamaan,

,

yaitu N adalah jumlah dokumen yang terdapat pada kelas c, N adalah total dokumen, 𝑇𝑐𝑡 adalah banyaknya token t dalam dokumen training dari kelas c (Manning *et al.* 2009). Nilai peluang P(t|c) dapat bernilai nol jika suatu kata dalam data uji tidak ada pada data latih. Untuk itu, digunakan Laplace Smoothing, dengan persaaman,

dengan *B* adalah jumlah *vocabulary* atau kata unik yang didapat dari data latih. Penambahan 1 dilakukan untuk mencegah terjadinya nilai peluang nol pada suatu kata. (Manning et al. 2008)

## Evaluasi

Kinerja hasil klasifikasi dapat dinilai berdasarkan beberapa pengukuran. Pada penelitian ini mengunakan teknik perhitungan *precision*, *recall* dan akurasi. Untuk memudahkan setiap perhitungan, maka digunakan tabel pendukung yaitu *Confussion Matrix*. *Confussion matrix* mengandung informasi tentang hasil aktual dan prediksi dari proses klasifikasi yang dilakukan oleh sistem (Han *et al*. 2012). Tabel 1 merupakan bentuk dari *confusion matrix*.

Tabel 1 *Confussion Matrix* untuk 3 kelas

|  |  |  |  |
| --- | --- | --- | --- |
| Aktual | Prediksi | | |
| Positif | Negatif | Netral |
| Positif | TP | FNg1 | FNt1 |
| Negatif | FP1 | TNg | FNt2 |
| Netral | FP2 | FNg2 | TNt |

Tabel 1 merupakan hasil informasi mengenai data hasil klasifikasi yang benar dan yang salah. Adapun dari Tabel 1 nilai *precison* dapat dihitung dengan persamaan ,

dengan TP (*True positive*) merupakan jumlah data hasil klasifikasi prediksi yang benar terhadap kelas aktual positif, FP1 (*False Positive* 1) merupakan jumlah data hasil prediksi positif terhadap kelas aktual negative, dan FP2 (*False Positive* 2) merupakan jumlah data hasil prediksi positif terhadap kelas aktual netral. Untuk nilai recall dapat diperoleh dengan,

yaitu FNg1(*False Negative* 1) merupakan jumlah data hasil prediksi negatif terhadap kelas aktual positif dan FNt1(*False Neutral* 1) merupakan jumlah data hasil prediksi neral terhadap kelas aktual positif, selanjutnya akurasi dapat diperoleh dengan persamaan,

dengan hasil pembagian antara semua hasil prediksi benar terhadap kelas aktualnya yaitu TP, TNg, dan TNt dengan jumlah semua data hasil prediksi benar dan salah.

***Clustering***

Menurut Han dan Kamber (2006) *clustering* merupakan proses pengelompokan objek ke dalam suatu kelas berdasarkan kesamaan sehingga objek yang terdapat dalam *cluster* tersebut memiliki kemiripan. Pada penelitian ini algoritma *k-means* yang digunakan untuk proses *clustering*, denga nilai k yang digunakan sebesar 3 yaitu untuk cluster positif, negatif, dan netral. Adapun pelabelan clusternya dengan cara mencari frekuensi terbanyak sentimen setiap cluster tersebut. Jika cluster dengan nilai sentimen tweet lebih banyak yang positif maka label cluster tersebut adalah positif, begitu juga dengan negatif dan netral. Proses *clustering* dilakukan untuk melihat pengelompokan data *tweet* yang digunakan. Data *tweet* yang telah di *clustering* akan mendapatkan pertimbangan yaitu sentiment ulang atau penghapusan *tweet*.

## Lingkungan Pengembangan

Spesifikasi perangkat keras dan perangkat lunak yang digunakan untuk penelitian ini adalah sebagai berikut:

1. Perangkat keras berupa komputer personal dengan spesifikasi sebagai berikut :
2. *Processor* AMD Phenom II X4 BE 3.2 GHZ
3. RAM 10 GB
4. *Harddisk* 2 TB HDD
5. VGA GTX 750 Ti
6. Perangkat lunak :
7. Sistem Operasi Windows 10 64-bit
8. PHP 5.6
9. DBMS MySQL
10. *Web server* Apache 3.2.2
11. Laravel 5.3

# HASIL DAN PEMBAHASAN

## Praproses

Data *tweet* yang didapatkan dari system *tags.hawskey.info* dengan keyword “IPB”, “Institut Pertanian Bogor” dan “Bogor Agricultural University” sebanyak 16635 data. Data tersebut diambil pada tanggal 10 September 2016 sampai dengan 16 September 2016. Data tersebut kemudian dibersihkan dengan menghapus tweet dengan Bahasa selain Bahasa Indonesia seperti tweet “عار ما يفعله عبدالعزيز محمد السعد العجلان الكهل و #أبناء\_محمد\_السعد\_العجلان بمحاربة ابناء عمهم وخالهم عجلان واخوانه Ipb” atau “Как продать автомобиль самостоятельно форум powered by ipb Автору киров ЖМИ НА ССЫЛКУ ➡ https://t.co/dUZIGt8AEC https://t.co/NOFDGdwA0t” ditemukan dalam data yang diperoleh maka tweet tersebut akan dihapus. Selanjutnya adalah penghapusan tweet yang duplikat. Dari kedua proses diatas data tweet terpotong sebanyak 13.721 data, sehingga menghasilkan 2914 data.

Dari data yang telah dibersihkan akan dilakukan pemberian label sentimen. Karena pemberian label sentiment dilakukan secara subyektif makan pemberian sentimen dilakukan oleh 5 orang. Dari setiap tweet akan disentimen oleh 5 orang dan hasil sentimen diambil dari sentimen yang muncul paling banyak. Hasil yang didapatkan adalah 512 kelas positif, 113 kelas negatif dan 2289 kelas netral.

Dalam penelitian ini data yang dipakai adalah data seimbang sehingga data akan dipotong menurut jumlah kelas sentimen yang terkecil yaitu kelas negatif yang berjumlah 113 tweet. Karena data yang didapatkan masih sedikit untuk proses pelatihan maka dilakukan pengumpulan data lagi dengan kata kunci “pilgub” dan “pilkadaDKI” pada tanggal 16 November 2016 sampai 17 November 2016 dan diperoleh sebanyak 204 data *tweet.* Data ditambahkan lagi dari penelitian Aziz (2013) sebanyak 27.275 data dan dari penelitian Adityawan (2014) sebanyak 1297 data. Jumlah data setiap kelas pada penelitian dapat dilihat pada Tabel 3.

Tabel Jumlah Tweet yang dihasilkan dari pembersihan data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sumber Data | Rincian jumlah *tweet* yang dihasilkan | | | |
| Jumlah tweet | Positif | Negatif | Netral |
| API Hawskey |  |  |  |  |
| IPB, Institut Pertanian Bogor dan *Bogor Agricultural University* | 2914 | 512 | 113 | 2289 |
| Pilkada dan pilkadaDKI | 204 | 60 | 44 | 100 |
| Data Penelitian |  |  |  |  |
| Aziz (2012) | 27275 | 1179 | 405 | 25691 |
| Adityawan (2014) | 1297 | 341 | 503 | 453 |
| Jumlah | 31690 | 2092 | 1065 | 28533 |

Dari Tabel 3 kelas dengan data paling kecil ada di kelas Negatif sebanyak 1065 data, maka data akan dipotong dipotong sesuai kelas Negatif sehingga setiap sentimen berjumlah 1065 data dan total data yang digunakan adalah 3195.

## *K-fold Cross Validation*

Data akan dibagi menjadi 10 subset data dengan sentimen imbang setiap subsetnya. Dari total data 3195 dengan 1065 data setiap sentimen, maka setiap subset akan memiliki 106 data positif, 106 data negative dan 106 data negatif. Total setiap subset adalah 318 data. Karena data memiliki sisa 15 data maka subset terakhir akan mempunyai 333 data. Setiap subset dalam penelitian ini akan disebut sebagai model 1 sampai model 10. Model 1 yaitu ketika subset pertama akan dijadikan data uji dan subset yang lain dijadikan data latih, Model 2 yaitu subset kedua dijadikan data uji dan sisanya akan dijadikan data latih begitu seterusnya sampai dengan model 10.

## Tokenisasi

Data yang berjumlah 3195 data tweet kemudian dilakukan tokenisasi dengan membuang karakter ASCII dengan nilai 32 sampai 126, url, angka, dan simbol lainnya dengan mengantinya dengan spasi menggunakan fungsi regex. Dalam proses tokenisasi didapatkan sebanyak 47381 token.

## Normalisasi Kata

Dalam proses normalisasi kata menghasilkan token sebanyak 47602 dan sebanyak 6566 token ternormalisasi. Dalam proses normalisasi kata terdapat penambahan token sebanyak 221 token karena dalam dataset normalisasi kata terdapat normalisasi kata dengan kata baku lebih dari satu kata seperti kata “yaudah” menjadi “ya sudah”. Dalam proses normalisasi terdapat kata yang belum masuk dataset normalisasi yaitu seperti kata “donk” akan diubah menjadi “dong” dan kata “koq” akan diubah menjadi “kok”. Kata yang belum masuk dalam daftar normalisasi kata Adityawan (2014) akan ditambahkan dalam dataset normalisasi. Daftar tambahan kata normalisasi dapat dilihat pada lampiran 1.

## Negation Handling

Data tweet yang telah melewati proses normalisasi kemudian masuk ke proses Negation Handling. Pada proses ini dihasilkan 591 tweet yang mengalami negasi dengan jumlah data tweet negasi yang mempunyai sentimen positif sebanyak 125, sentimen negatif sebanyak 303 dan sentimen netral sebanyak 161. Jumlah token negasi yang dihasilkan dalam proses Negation Handling adalah sejumlah 46988 token, dengan jumlah token sebanyak 15482 dalam token positif, 16734 dalam token negatif, 14772 dalam token netral.

Adapun beberapa kata yang mengalami negation handling, seperti pada potongan data yang ditampilkan pada Tabel 3.

Tabel 3. 10 kata negasi dengan *term frequency* (TF) terbanyak

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kata | TF Positif | TF Negatif | TF Netral | TF |
| tidak\_bisa | 11 | 125 | 18 | 154 |
| tidak\_ada | 12 | 33 | 19 | 64 |
| tidak\_mau | 0 | 11 | 4 | 15 |
| tidak\_di | 2 | 9 | 2 | 13 |
| tidak\_hanya | 5 | 0 | 8 | 13 |
| tidak\_stabil | 9 | 1 | 2 | 12 |
| tidak\_pernah | 1 | 2 | 6 | 9 |
| tidak\_perlu | 3 | 2 | 4 | 9 |
| tidak\_akan | 3 | 2 | 3 | 8 |
| tidak\_ya | 2 | 3 | 3 | 8 |

Dalam proses Negation Handling yang dilakukan Narayanan (2013) untuk mencukupi kebutuhan kata negasi dalam klasifikasi, maka setiap token dalam data latih akan mengalami negasi menjadi bentuk “tidak\_” + kata. Dengan penambahan token tersebut maka jumlah token didapatkan melalui proses ini berjumlah 61760 token.

Dalam 10 model percobaan yang dilakukan, distribusi *tweet* dalam data latih dan data uji dapat dilihat pada Tabel ..

Tabel 4 Distribusi *tweet* negasi pada data latih dan data uji dengan 10-fold Cross Validation

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Data latih | Data uji | |
| 1 | 546 | 45 |
| 2 | 547 | 44 |
| 3 | 532 | 59 |
| 4 | 525 | 66 |
| 5 | 515 | 76 |
| 6 | 526 | 65 |
| 7 | 526 | 65 |
| 8 | 510 | 81 |
| 9 | 525 | 66 |
| 10 | 553 | 38 |

## Pembuangan Stopword

Proses pembuangan *stopword* menggunakan dataset sebanyak 763 kata dan terdapat sebanyak 15556 kata yang dihapus pada tahap ini. Dalam token yang dihasilkan pada penelitian ini ada kata seperti “x”, “http”, “amp”, “at”, “rt” yang tidak memiliki informasi untuk klasifikasi, maka kata tersebut ditambahkan kedalam daftar kata stopword. Daftar *stopword* yang ditambahkan ada pada lampiran 2. Dari seluruh rangkaian proses yang dilakukan jumlah token yang dihasilkan dapat dilihat pada tabel 4.

Tabel Jumlah token yang dihasilkan dari setiap tahapan

|  |  |
| --- | --- |
| Tahapan | Jumlah token yang dihasilkan |
| Tokenisasi | 47381 |
| Normalisasi | 47602 |
| *Negation Handling* | 61760 |
| Penghapusan *Stopword* | 30648 |

Token yang dihasilkan disimpan sebagai kata unik dihitung nilai tf dan df nya. Kata unik yang dihasilkan dari rangkain proses diatas adalah sebanyak 16124. Tabel 10 merupakan 10 daftar kata unik beserta nilai tf dan df dengan tf terbanyak.

Tabel 10 Sepuluh daftar kata unik beserta nilai tf dan df dengan tf terbanyak

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Kata Unik | TF | TF | | | DF |
| Positif | Negatif | Netral |
| tidak\_ipb | 867 | 156 | 346 | 365 | 828 |
| ipb | 867 | 346 | 156 | 365 | 828 |
| tidak\_ya | 439 | 157 | 89 | 193 | 415 |
| ya | 439 | 89 | 157 | 193 | 415 |
| tidak\_xl | 425 | 270 | 86 | 69 | 363 |
| xl | 425 | 86 | 270 | 69 | 363 |
| tidak\_xlcare | 402 | 274 | 69 | 59 | 402 |
| xlcare | 402 | 69 | 274 | 59 | 402 |
| tidak\_terimakasih | 398 | 29 | 291 | 78 | 397 |
| terimakasih | 398 | 291 | 29 | 78 | 397 |

## Seleksi fitur

Kata unik yang sudah didapatkan dihitung nilai idf nya. Rentang nilai idf yang didapatkan adalah 0.586 sampai 3.504. Pada proses pemilihan ini akan dicoba memotong kata unik dari nilai idf yang paling tinggi hingga paling rendah dan dihitung nilai akurasi nya. Pemotongan nilai idf dengan akurasi yang paling tinggi akan dipakai dalam memilih fitur yang akan digunakan. Pada gambar 2 adalah gambar jumlah fitur yang terpilih berikut dengan akurasinya.

Gambar 2 Jumlah fitur yang terpilih dengan akurasinya.

Pada gambar 1 pemotongan yang dilakukan menimbulkan penurunan akurasi karena fitur berkurang. Dalam penelitian ini akan digunakan pemotongan dengan nilai idf lebih dari 3.2 sehingga kata dengan nilai df kurang dari 3 akan dihapus. Pemotongan dengan nilai idf lebih dari 3.2 menurunkan akurasi sebesar 1.47% namun meningkatkan waktu komputasi sebesar 58% dari rata-rata waktu komputasi sebesar 104.05 detik menjadi 27.584 detik.

## Evaluasi

Model yang digunakan adalah 10-fold cross validation. Dari 10 model yang dilakukan evaluasi, akurasi terbaik didapatkan dari model 9 dengan akurasi sebesar 79.25% tanpa menggunakan *Negation Handling* dam dengan menggunakan *Negation Handling* akurasi menurun menjadi sebesar 78.93%/. *Confusion matrix* dari model 9 tanpa menggunakan *Negation Handling* dapat dilihat pada Tabel 7 dan dengan menggunakan *Negation Handling* pada Tabel 8.

Tabel 7 *Confusion matriks* hasil klasifikasi *Naïve Bayes* tanpa *Negation Handling*

|  |  |  |  |
| --- | --- | --- | --- |
| Fakta | Prediksi | | |
| Positif | Negatif | Netral |
| Positif | 77 | 8 | 7 |
| Negatif | 8 | 85 | 9 |
| Netral | 21 | 13 | 90 |

Tabel 8 *Confusion matriks* hasil klasifikasi *Naïve Bayes* menggunakan *Negation Handling*

|  |  |  |  |
| --- | --- | --- | --- |
| Fakta | Prediksi | | |
| Positif | Negatif | Netral |
| Positif | 75 | 8 | 5 |
| Negatif | 9 | 85 | 10 |
| Netral | 22 | 13 | 91 |

Dari Tabel *Confusion matrix* tanpa menggunakan *Negation Handling* hasil klasifikasi yang salah berjumlah 66 dari 318 data, dengan menggunakan *Negation Handling* hasil klasifikasi yang salah berjumlah 67. Hasil *precision* dan *recall* dari model 7 dapat dilihat pada Tabel 9.

Tabel Nilai *precision* dan *recall* dari model 7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Klasifikasi | Precision | | | Recall | | |
| Positif | Negatif | Netral | Positif | Negatif | Netral |
| Multinomial NB | 68.62 | 74.599 | 63.168 | 61.721 | 71.371 | 70.51 |
| Multinomial NB  + *Negation Handling* | 69.13 | 74.36 | 63.28 | 60.86 | 72.69 | 70.70 |

Akurasi rata-rata yang didapatkan dari klasifikasi tanpa menggunakan *Negation Handling* adalah 67.87 sedangkan dengan menggunakan *Negation Handling* akurasi menjadi 68.08. Nilai akurasi dari 10 model yang dicobakan dapat dilihat pada Tabel 11.

Tabel  Akurasi Hasil Klasifikasi *Multinomial Naive Bayes* tanpa menggunakan *Negation Handling* dan menggunakan *Negation Handling*

|  |  |  |
| --- | --- | --- |
| Model ke- | Akurasi Klasifikasi(%) | |
| *Multinomial* NB | *Multinomial* NB + *Negation Handling* |
| 1 | 55.66 | 56.92 |
| 2 | 76.42 | 77.99 |
| 3 | 75.79 | 77.99 |
| 4 | 70.75 | 69.50 |
| 5 | 66.04 | 66.67 |
| 6 | 61.95 | 61.64 |
| 7 | 57.86 | 57.86 |
| 8 | 69.81 | 67.92 |
| 9 | 79.25 | 78.93 |
| 10 | 65.17 | 65.47 |
| Rata-rata | 67.87 | 68.09 |

Dari Tabel 11 didapatkan rata-rata dengan menggunakan *Negation Handling* lebih tinggi daripada tanpa menggunakan *Negation Handling* yaitu sebesar 0.22%. Walaupun rata-rata menunjukkan peningkatan akurasi namun jika dilihat setiap model 4, 6, 8 dan 9 mengalami penurunan akurasi. Penurunan akurasi paling banyak terjadi pada model 8 dengan penurunan sebesar 1.89%. Tweet dengan klasifikasi berbeda tanpa menggunakan *Negation Handling* dan menggunakan *Negation Handling* dapat dilihat pada Tabel 12.

Terdapat 6 tweet yang membedakan klasifikasi tanpa menggunakan *Negation handling* dan menggunkan *Negation Handling.* Tweet seperti “Ilmupemerintah ngaku kartikadjoemadi korupsi jokowi walikota solo” mendapatkan klasifikasi salah jika menggunakan *Negation Handling* dari tweet yang mempunyai sentimen positif hasil klasifikasinya menjadi negatif. Nilai tf dari tweet tersebut dapat dilihat pada Tabel 13 dan Tabel 14.

Tabel 12 Tweet dengan klasifikasi berbeda tanpa menggunakan *Negation Handling* dan menggunakan *Negation Handling*

|  |  |  |  |
| --- | --- | --- | --- |
| Tweet | Sentimen | | Sentimen tweet |
| Tanpa Negation Handling | Negation Handling |
| Ilmupemerintah ngaku kartikadjoemadi korupsi jokowi walikota solo | positif | negatif | positif |
| giofedi iya pas wiranto diduga korupsi busway bekas senilai muak banged sebastian dewadino lordrio luviku | negatif | netral | negatif |
| danvisarah dear susipudjiastuti anak ipb banget kerja sektor agribisnis bumn ya buka ya saran | negatif | netral | negatif |
| panas ihh etapi panasan lihat kampus ipb gunung | negatif | positif | negatif |
| min register ya register xlcare naruchan jrs pagi mas naruchan dibantu layanan | positif | negatif | positif |
| follow back dm nomor pascabayar thxrt xlcare hannymadia hai mbak hanny idbantu layanan xlnya ds | netral | positif | netral |

Tabel 13 Nilai tf tweet tanpa menggunakan *Negation Handling*

|  |  |  |  |
| --- | --- | --- | --- |
| Kata | tf | | |
| positif | negatif | netral |
| ilmupemerintah | 0 | 0 | 0 |
| ngaku | 1 | 1 | 2 |
| kartikadjoemadi | 0 | 0 | 0 |
| korupsi | 4 | 2 | 0 |
| jokowi | 33 | 58 | 7 |
| walikota | 0 | 0 | 0 |
| solo | 0 | 0 | 0 |

Tabel 14 Nilai tf tweet menggunakan *Negation Handling*

|  |  |  |  |
| --- | --- | --- | --- |
| Kata | tf | | |
| positif | negatif | netral |
| ilmupemerintah | 0 | 0 | 0 |
| ngaku | 1 | 0 | 2 |
| kartikadjoemadi | 0 | 0 | 0 |
| korupsi | 0 | 0 | 0 |
| jokowi | 33 | 58 | 7 |
| walikota | 0 | 0 | 0 |
| solo | 0 | 0 | 0 |

Dari Tabel 13 dan 14 dapat dilihat perbedaan terletak pada kata “ngaku” dan “korupsi”. Pada kata “ngaku” tanpa menggunakan *Negation Handling* nilai tf dalam kelas negatif berjumlah 1 sedangkan dengan menggunakan *Negation Handling* nilai nya berubah menjadi 0. Dalam kata “korupsi” juga terdapat perbedaan dalam kelas potifif bernilai 4 menjadi 0, dalam kelas negative bernilai 2 menjadi 0. Hal tersebut dikarenakan kata “ngaku” dan “korupsi” menjadi kata negasi “tidak\_ngaku” dan “tidak\_korupsi” sehingga kata tersebut tidak ditemukan ketika menggunakan *Negation Handling* sehingga mengakibatkan klasifikasi yang salah.

Untuk menangani masalah diatas, dalam penelitian ini akan ditambahkan fitur negasi dari kata negasi, seperti kata yang sudah mengalami negasi yaitu “tidak\_baik” yang mempunyai tf positif 10 dan tf negatif 120 akan dinegasikan kembali menjadi kata “baik” dengan nilai tf kebalikan dari kata tersebut menjadi tf positif 120 dan negatif 10. Hasil akurasi dari penambahan fitur ini dapat dilihat pada Tabel 15.

Tabel 5 Akurasi Hasil Klasifikasi *Multinomial Naive Bayes* tanpa menggunakan *Negation Handling* dan menggunakan *Negation Handling*

|  |  |  |
| --- | --- | --- |
| Model ke- | Akurasi Klasifikasi(%) | |
| *Multinomial* NB | *Multinomial* NB + *Negation Handling* |
| 1 | 55.66 | 65.77 |
| 2 | 76.42 | 78.93 |
| 3 | 75.79 | 68.87 |
| 4 | 70.75 | 57.23 |
| 5 | 66.04 | 61.01 |
| 6 | 61.95 | 66.35 |
| 7 | 57.86 | 70.13 |
| 8 | 69.81 | 78.3 |
| 9 | 79.25 | 77.99 |
| 10 | 65.17 | 57.55 |
| Rata-rata | 67.87 | 68.21 |

Dari Tabel 13 hasil akurasi mengalami peningkatan sebesar 0.34% dengan menggunakan *Negation Handling* yang ditambahkan fitur dalam penelitian ini. Dengan penambahan fitur tersebut terdapat peningkatan sebsesar 0.12% dibandingkan dengan *Negation Handling* dari penelitian Narayanan (2014).

## Clustering

Hasil akurasi dari klasifikasi belum menunjukan hasil yang baik, untuk itu dilakukan proses clustering dengan menggunakan k-means untuk melihat pengelompokan datanya. Pengelompokan data tweet ini bertujuan untuk dijadikan pertimbangan apakah terdapat kesalahan dalam melakukan pelabelan tweet secara manual. Setelah dilakukan clustering kita akan mendapatkan data setiap cluster dan jumlah tweet sentiment positif, negatif, dan netral pada tiap cluster tersebut. Adapun hasil clustering yang didapatkan yaitu seperti yang ditampilkan pada tabel 16.

Tabel 6 Confusion matriks hasil clustering

|  |  |  |  |
| --- | --- | --- | --- |
| Hasil Kmeans | Sentimen awal | | |
| Positif | Negatif | Netral |
| Positif | 293 | 275 | 182 |
| Negatif | 566 | 617 | 410 |
| Netral | 206 | 173 | 473 |

Dari Tabel 16 didapatkan sebanyak 1812 tweet yang mempunyai label *cluster* berbeda dengan sentimen tweet. Tweet dengan label *cluster* berbeda dengan sentimen tweet dilakukan pemberian sentiment ulang dan mendapatkan 64 tweet mengalami perubahan sentimen. Perubahan sentimen tersebut dapat dilihat pada Tabel 17.

Tabel 7 Perubahan sentimen tweet

|  |  |  |  |
| --- | --- | --- | --- |
| Sentimen awal | Sentimen setelah perubahan | | |
| Positif | Negatif | Netral |
| Positif | 0 | 16 | 12 |
| Negatif | 7 | 0 | 5 |
| Netral | 9 | 14 | 0 |

Dengan perubahan sentimen tersebut maka *confusion matrix* hasil *clustering* berubah sesuai dengan tweet yang berubah sentimennya. *Confusion matrix* setelah dilakukan perubahan sentimen dapat dilihat pada Tabel 18.

Tabel 8 *Confusion matriks* hasil clustering setelah perubahan setimen tweet

|  |  |  |  |
| --- | --- | --- | --- |
| Hasil Kmeans | Sentimen awal | | |
| Positif | Negatif | Netral |
| Positif | 310 | 269 | 171 |
| Negatif | 548 | 643 | 402 |
| Netral | 196 | 170 | 486 |

Langkah selanjutnya adalah menghapus *tweet* dengan label *cluster* berbeda dengan sentimen tweet, sehingga jumlah data *tweet* yang dipakai menjadi 310 untuk kelas positif, 643 untuk kelas negatif, dan 486 kelas netral. Hasil dari proses *clustering* kemudian dilakukan evaluasi kembali dan didapatkan nilai akurasi seperti yang ditampilkan pada tabel 15.

Tabel 5 Akurasi Hasil Klasifikasi setelah proses *clustering*

|  |  |  |
| --- | --- | --- |
| Model ke- | Akurasi Klasifikasi(%) | |
| *Multinomial* NB | *Multinomial* NB + *Negation Handling* |
| 1 | 72.73 | 88.11 |
| 2 | 88.11 | 88.81 |
| 3 | 88.11 | 93.01 |
| 4 | 86.01 | 95.10 |
| 5 | 91.61 | 92.31 |
| 6 | 88.81 | 90.21 |
| 7 | 89.51 | 93.71 |
| 8 | 88.81 | 91.61 |
| 9 | 95.10 | 94.41 |
| 10 | 92.76 | 88.82 |
| Rata-rata | 88.16 | 91.61 |

Hasil klasifikasi setelah proses *clustering* meningkatkan akurasi *Multinomial Naïve Bayes* sebesar 20.29% sedangkan untuk *Multinomial Naïve Bayes* dengan penambahan *Negation Handling* menigkat sebesar 23.4%. Hasil akurasi yang didapatkan menunjukkan peningkatan akurasi sebesar 3.45% dengan menggunakan *Negation Handling* yang ditambahankan fitur negasi dari kata yang telah mengalami negasi dalam penelitian ini. Penambahan fitur tersebut juga meningkatkan akurasi sebesar 0.14% dibandingkan metode *Negation Handling* yang digunakan Narayanan (2014).

Model dengan peningkatan paling banyak adalah model 1 dengan peningkatan sebesar 15%. Rata-rata p*recision dan recall* data setelah clustering dapat dilihat pada tabel 16.

Tabel Nilai rata-rata *precision* dan *recall* dari data setelah *clustering*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Klasifikasi | Precision | | | Recall | | |
| Positif | Negatif | Netral | Positif | Negatif | Netral |
| Multinomial NB | 70.00 | 98.87 | 95.157 | 98.79 | 89.11 | 80.09 |
| Multinomial NB  + *Negation Handling* | 100 | 94.03 | 79.37 | 70.97 | 94.03 | 92.59 |

# KESIMPULAN DAN SARAN

## Kesimpulan

Berdasarkan pernelitian yang telah dilakukan, dapat disimpulkan beberapa hal yaitu :

1. Penambahan fitur *Negation Handling* dapat diterapkan dengan menggunakan *multinomial naïve bayes* dan dengan data berbahasa Indonesia.
2. Dengan menggunakan *Negation Handling* sepeti dalam penelitian Narayanan (2014) dapat meningkatkan akurasi sebesar 3.45%.
3. Dengan menggunakan *Negation Handling* dalam penelitian dapat meningkatkan akurasi sebesar 3.59%, miningkat sebesar 0.14%.
4. *Clustering* dengan menggunakan *k-means* dapat digunakan untuk mengetahui pelabelan sentimen yang salah.
5. Dengan menghilankang data yang mempunyai sentimenyang berbeda dengan label *cluster* dapat meningkatkan akurasi sebesar 20.29% untuk klasifikasi *Multinomial Naïve Bayes* dan 23.4% untuk *Multinomial Naïve Bayes* ditambahkan dengan *Negation Handling.*

## Saran

Dalam penelitian selanjutnya dapat dilakukan dengan menggunakan data yang lebih banyak dan juga mempunyai tweet negasi yang lebih banyak sehingga meningkatkan peluang dari kata negasi.

# DAFTAR PUSTAKA

Adityawan, E. 2014. “Analisis Sentimen Dengan Klasifikasi Na¨ıve Bayes Pada Pesan Twitter Menggunakan Data Seimbang”. Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor.

Adriani, Mirna et al. 2007. “Stemming Indonesian: A Confix-stripping Approach”. 6 (4), pp. 1–33. ISSN: 1530-0226. DOI: 10.1145/1316457.1316459.

Agusta, Ledy. 2009. “Perbandingan Algoritma Stemming Porter dengan algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia”. Universitas Kristen Satya Wacana.

Alkhatib, K, Najadat H, Hmeidi I, dan Shatnawi MKA. 2013. “Stock Price Prediction Using k-Nearest Neighbor (kNN) Algorithm”.

Anwar Hridoy, Syed Akib et al. 2015. “Localized twitter opinion mining using sentiment analysis” dalam*: Decision Analytics* 2 (1), pp. 1–19. ISSN: 2193-8636. DOI: 10.1186/s40165-015-0016-4.

Aziz, ATA. 2013. “Sistem pengklasifikasian entitas pada pesan twitter menggunakan ekspresi reguler dan na¨ıve Bayes”. Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor.

Das, Sanjiv R. et al. 2001. “Yahoo! for amazon: Sentiment extraction from small talk on the web”. *8th Asia Pacific Finance Association Annual Conference*.

Dimastyo, JG. 2014. “Pengukuran Kinerja Spam Filter dengan Seleksi Fitur yang berbeda menggunakan Fungsi Klasifikasi Multinomial Na¨ıve Bayes”. Skripsi. Departemen Ilmu Komputer, Institut Pertanian Bogor.

Feinerer, Ingo, Kurt Hornik, dan David Meyer. 2008. “Text Mining Infrastructure in R” dalam: *Journal of Statistical Software* 25 (5), pp. 1–54.

Garnes. 2009. “Feature Selection for Text Categorisation”. Norwegia (NO): NTNU.

Liu, Bing. 2010. “Sentiment Analysis and Subjectivity, in Handbook of Natural Language Processing”. Chicago (US): University of Illinois.

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan dan Claypool Publishers. [Internet]. [Diunduh tanggal 11/08/2016 ]. Dapat diunduh dari: https: //www.cs.uic.edu/˜liub/FBS/SentimentAnalysisand-OpinionMining.pdf.

Manning, Christopher D., Prabhakar Raghavan, dan Hinrich Sch¨utze. 2008. *An Introduction to Information Retrieval.* Cambridge University Press Cambridge,

England.

Mudinas, Andrius, Dell Zhang, dan Mark Levene. 2012. “Combining Lexicon and Learning Based Approaches for Concept-level Sentiment Analysis”. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. WISDOM ’12. Beijing, China: ACM, 5:1–5:8. ISBN: 978-1-4503-1543-2. DOI: 10.1145/2346676.2346681.

Nadilah. 2016. “Asosiasi dan Geovisualisasi Antara Data Tweet Terkait Kebakaran Hutan dengan Data Cuaca di Provinsi Riau dan Kepulauan Riau”. Institut Pertanian Bogor.

Narayanan, V, I Arora, dan A Bhatia. 2013. “Fast and accurate sentiment classification using an enhanced Naive Bayes model”. Department of Electronics Engineering, Indian Institute of Technology (BHU), Varanasi, India.

Pak, Alexander dan Patrick Paroubek. 2010. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Ed. by Nicoletta Calzolari (Conference Chair) *et al*. Valletta, Malta: European Language Resources Association (ELRA). ISBN: 2-9517408-6-7.

Pang, Bo, Lillian Lee, dan Shivakumar Vaithyanathan. 2002. “Thumbs Up? Sentiment Classification Using Machine Learning Techniques”. *Proceedings of EMNLP*, pp. 79–86.

Robertson, Stephen. 2004. “Understanding inverse document frequency: On theoretical arguments for IDF” dalam: *Journal of Documentation* 60, p. 2004.

Routray, Preeti, Chinmaya Kumar Swain, dan Smita Praya Mishra. 2013. “Article: A Survey on Sentiment Analysis” dalam: *International Journal of Computer Applications* 76 (10). Full text available, pp. 1–8.

Seiberg, Nathan dan Edward Witten. 1999. “String theory and noncommutative geometry” dalam*: Journal of High Energy Physics* 1999 (09), p. 032.

Sproat, Richard et al. 2001. “Normalization of nonstandard words.” dalam: *Computer Speech & Language* 15 (3), pp. 287–333.

Tala, Fadillah Z. 2003. “A study of stemming effects on information retrieval in Bahasa Indonesia” dalam: *Institute for Logic, Language and Computation Universeit Van Amsterdam*.

Xhemali, Daniela, Christopher J. HINDE, dan Roger G. STONE. 2009. *Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training*

*Web Pages*.

Zhang et al. 2011. “Combining Lexicon-based and Learningbased Methods for Twitter Sentiment Analysis”.