PERANCANGAN KLASIFIKASI TWEET BERDASARKAN SENTIMEN DAN FITUR CALON GUBERNUR DKI JAKARTA 2017

Sumarni Adi

Informatika
Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Depok, DIY 55281, Indonesia

sumarni.a@amikom.ac.id

Abstract

One of the fastest growing social media users is Twitter, the number of twitter users mentioned continues to increase 300,000 users every day [1]. Twitter users send twitter posts about the facts and opinions of the government products or services they use or express their political, ideological and interest views. Not to mention also send tweet opinions related leaders or influential public figures in this country. With 55 million tweets each day Twitter has a high update rate [1] and is a highly efficient data warehouse for political and social research, so Twitter is a good place to conduct opinion mining or sentiment analysis in classifying the 2017 Jakarta governor candidate .

The classification of tweet data is done by analyzing the sentiments on Indonesian tweet opinions by extracting features using Unigram, negation, term Frequency, and TF-IDF (Term Frequency-Invers Document Frequency). Once extracted, the tweet is classified using the Naïve Bayes Classifier (NBC) algorithm.

From the results of designing the twitter classification of Indonesian language using Naïve Bayes Classifier algorithm obtained significant difference in value when compared with manual labeling. Positive and neutral sentiments are significant, while negative sentiments are not significant.

Keywords: tweet, sentiment, classification, Naïve Bayes Classifier (NBC)

Abstrak

Salah satu media sosial yang berkembang sangat pesat penggunanya adalah Twitter, jumlah pengguna twitter disebutkan terus meningkat 300.000 user setiap harinya[1]. Para pengguna twitter mengirimkan twitter post (tweet) mengenai fakta dan opini produk atau layanan pemerintah yang mereka gunakan atau mengekspresikan pandangan politik, ideologis dan minat mereka. Tidak terkecuali juga mengirimkan tweet opini terkait pemimpin atau tokoh publik yang berpengaruh di negara ini. Dengan 55 juta tweet setiap harinya menjadikan twitter memiliki tingkat update yang tinggi[1] dan menjadi gudang data yang sangat efisien untuk penelitian dibidang politik dan sosial, sehingga twitter merupakan tempat yang baik untuk melakukan opinion mining atau analisis sentimen dalam mengklasifikasikan calon gubernur DKI Jakarta 2017.

Pengklasifikasian data tweet dilakukan dengan cara melakukan analisis sentimen pada opini tweet berbahasa indonesia dengan mengekstraksi fitur menggunakan Unigram, negation, term Frequency, dan TF-IDF (Term Frequency-Invers Document Frequency). Setelah diekstraksi, tweet tersebut diklasifikasi menggunakan algoritma Naïve Bayes Classifier (NBC).

Dari hasil perancangan klasifikasi twitter berbahasa Indonesia menggunakan algoritma Naïve Bayes Classifier didapatkan perbedaan nilai yang cukup signifikan bila dibandingkan dengan pelabelan manual. Sentiment positif dan netral cukup signifikan perbedaannya, sedangkan sentiment yang bernilai negatif tidak terlalu signifikan.

Kata Kunci : *tweet, sentimen, klasifikasi, naïve bayes classifier(NBC)*

I. PENDAHULUAN

Mikroblog seperti Twitter (www.twitter.com) dan Facebook (www.facebook.com)sekarang menjadi perangkat komunikasi yang sangat populer dikalangan pengguna internet. Salah satu media sosial yang berkembang sangat pesat penggunanya adalah Twitter, jumlah pengguna twitter disebutkan terus meningkat 300.000 user setiap harinya[1]. Pada konferensi resmi pengembang twitter Chirp 2010, perusahaan tersebut menyampaikan statistik mengenai situs dan penggunaan twitter. Statistik tersebut menyebutkan bahwa pada bulan april 2010, twitter memiliki 106 juta akun dan sebanyak 180 juta pengunjuk unik setiap bulannya. Jumlah pengguna twitter disebutkan terus meningkat 300.000 user setiap harinya[1]. Setiap harinya para pengguna twitter mengirimkan twitter post (tweet) mengenai fakta dan opini produk atau layanan pemerintah yang mereka gunakan atau mengekspresikan pandangan politik, ideologis dan minat mereka. Tidak terkecuali juga mengirimkan

tweet opini terkait pemimpin atau tokoh publik yang berpengaruh di negara ini. menyediakan sumber-sumber opini yang besar jumlahnya bagi kebutuhan individu maupun organisasi. Melalui media sosial orang dapat mengekspresikan apa saja, termasuk pendapatnya akan sesuatu hal tanpa adanya keterpaksaan.

Dengan 55 juta tweet setiap harinya[1] menjadikan twitter memiliki tingkat update yang tinggi. Hal inilah yang mengakibatkan tingginya ketersediaan data di twitter, sehingga twitter merupakan tempat yang baik untuk melakukan opinion mining atau analisis sentimen. Gudang data yang ada di twitter inilah yang sangat efisien untuk penelitian dibidang politik, pemasaran dan sosial.

Secara umum terdapat dua tipe informasi tekstual di web yaitu fakta dan opini. Fakta adalah pernyataan objektif mengenai entitas dan kejadian di dunia sedangkan opini adalah pernyataan subjektif yang merefleksikan sentimen atau persepsi orang mengenai entitas ataupun kejadian di dunia. Opini akan menjadi penting ketika calon Gubernur DKI Jakarta Tahun 2017 ingin membuat sebuah keputusan dengan terlebih dahulu mendengar opini dari pihak lain.Sekarang ketika calon gubernur DKI Jakarta tahun 2017 ingin memperoleh opini publik mengenai produk, citra dan layanannya, maka mereka tidak perlu melakukan survey konvensional dan focus group yang lama dan mahal biayanya.

Di beberapa Negara, Twitter telah dimanfaatkan untuk menjaring pendapat masyarakat terhadaptokoh publik dan juga prediksi calon legislatif yang akan terpilih seperti Singapura, Jerman dan Amerika [2]. Hal itu dikarenakan twitter merupakan salah satu media jejaring sosial dengan pengguna terbanyak diantara beberapa situs jejaring sosial yang ada. Hal ini juga menimbulkan motivasi khusus untuk mengembangkan suatu penelitian. Penelitian ini membahas tentang bagaimana melakukan klasifikasi dengan menganalisis sentimen pada opini *tweet* berbahasa indonesia pada calon Gubernur DKI Jakarta tahun 2017 dengan mengekstraksi fitur menggunakan Unigram, negation, term Frequency, dan TF-IDF (Term Frequency-Invers Document Frequency). Algoritma untuk melakukan klasifikasi menggunakan algoritma *Naive Bayes Classifier* (SVM).

II. TEORI

a. Analisis Sentimen/opinion Mining

Analisis sentimen/opinion mining merupakan sebuah cabang penelitian di domain text mining yang mulai marak pada tahun 2003. Opinion mining atau sentiment analysis adalah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual. Jika diberikan satu set dokumen teks D yang berisi opini (atau sentimen) mengenai suatu objek, maka opinion mining bertujuan untuk mengekstrak atribut dan komponen dari objek yang telah dikomentari pada setiap dokumen d ∈ D dan untuk menentukan apakah komentar tersebut positif atau negatif [3]. Opinion mining adalah bagian pekerjaan yang melakukan review yang berkaitan dengan perlakuan komputasional opini, sentimen dan subjektifitas dari teks [4].

b. Model Opinion Mining

Secara umum, opini dapat diekspresikan atas apa saja, misalnya produk, layanan, individu, organisasi, atau suatu kejadian. *Term object* digunakan untuk menunjukkan entitas yang telah dikomentari. Suatu objek memiliki seperangkat komponen dan satu set atribut [3]. Mendefinisikan bahwa suatu kalimat opini merupakan kalimat yang mengekspresikan opini positif atau negatif secara eksplisit atau implisit[3]. Suatu kalimat opini dapat berupa kalimat subjektif atau kalimat objektif[3]. Opini eksplisit merupakan opini yang secara eksplisit diekspresikan terhadap fitur atau objek dalam suatu kalimat subjektif. Sedangkan opini implisit merupakan opini terhadap fitur atau objek yang tersirat dalam suatu kalimat objektif [3]. Misalnya kalimat "kualitas suara dari telepon ini luar biasa" merupakan opini yang positif dan eksplisit. Sedangkan kalimat "earphone ini rusak dalam dua hari" merupakan opini yang negatif dan implisit. Meskipun kalimat "earphone ini rusak dalam dua hari" menyampaikan fakta objektif, namun secara implisit kalimat ini mengindikasikan opini negatif terhadap "earphone"[3]. Secara umum, kalimat objektif menyiratkan opini positif, negatif maupun netral.

c. Opini dan Orientasi Opini (Sentimen)

Suatu opini atas suatu feature f dari object adalah suatu pandangan positif, negatif, sikap, emosi atau penilaian pada f dari opinion holder [3]. Definisi opinion holder sebagai orang atau organisasi yang mengekspresikan suatu opini[3]. Opinion holder disebut juga opinion source. Orientasi opini berkaitan dengan istilah *polarity*. Orientasi opini atas suatu feature f dari object mengindikasikan apakah opini tersebut positif, ataukah negatif dapat diukur berdasarkan skala yang lebih granular untuk mengekspresikan perbedaan makna dari opini[3]. Skala yang lebih granular dari opini misalnya : sangat positif, positif, sangat negatif, dan negatif. Skala yang lebih granular ini biasanya melibatkan semantik dan sangat baik jika menggunakan POS-Tagger.

d. EkstraksiFitur

Ekstraksi fitur merupakan sebuah proses dimana properti-properti diekstrak dari suatu data [5]. Ekstraksi fitur dilakukan dengan tujuan untuk mengidentifikasikan sesuatu yang dimaksud. Adapun fitur-fitur yang rencana akan digunakan dalam penelitian ini adalah : Unigram, negation, term Frequency, dan TF-IDF (Term Frequency-Invers Document Frequency).

1. Unigram

Unigram feature extractor merupakan cara paling sederhana dalam mendapatkan fitur dari tweet[6]. Proses ekstraksi unigram dilakukan dengan mengekstrak kata per kata dalam dokumen. Unigram tidak memperhatikan konteks dan melakukan estimasi terhadap masing-masing term secara independen tanpa adanya keterkaitan suatu kata dengan kata lain [5]. Model unigram digambarkan dengan Persamaan 1.

$P_{uni}(t_1t_2t_3t_4) = P(t_1)\ P(t_2)\ P(t_3)\ P(t_4)$ .....(1)

2. Negation

Negasi merupakan sesuatu yang dikenal dalam semua bahasa dan biasanya negasi digunakan untuk mengubah polaritas dari suatu pernyataan [7]. Keberadaan negasi dalam sebuah kalimat menjadi sangat penting dikarenakan kemunculannya sering mengubah orientasi dari suatu opini [3]. Kata negasi dalam bahasa indonesia diantaranya kata tidak, tak, bukan, jangan, dan sebagainya. Kata-kata tersebut apabila diikuti oleh kata yang bernilai positif maka akan mengubah polaritas susunan kata menjadi negatif, begitu pula sebaliknya. Misalnya kata 'baik' merupakan kata yang memiliki polaritas sentimen positif. Jika kata "baik" didahului kata 'tidak' maka susunan kata menjadi 'tidak baik' yang berarti susunan kata tersebut bernilai negatif.

3. Term Frequency

TermFrequency ($tf_{t,d}$) yaitu jumlah kemunculan *term* t dalam suatu koleksi dokumen d [5]. Untuk menghindari nilai 0 pada probabilitas kondisional dengan menggunakan Metode Naive bayes maka dilakukan *smooting* yang disebut Laplace *smooting.*

4. TF-IDF

Pada dasarnya tf-idf bekerja dengan menghitung frekuensi relatif dari suatu kata yang muncul pada sebuah dokumen dibandingkan dengan proporsi invers dari kata tersebut yang muncul pada seluruh kumpulan dokumen [5]. Secara intuitif, perhitungan ini dapat digunakan untuk mengetahui seberapa relevan kata tersebut pada sebuah dokumen tertentu. Adapun rumus untuk menghitung tf-idf dapat dilihat di Persamaan 2.

$tfidf\ (d, w) = tf\ (d, w) \times \log N/dfw$ …….. (2)

Dimana : tf(d,w) adalah frekuensi kemunculan term w pada dokumen d, n adalah jumlah keseluruhan dokumen dan dfw adalah jumlah dokumen yang mengandung term w.

e. Naive Bayes Classifier

Salah satu metode klasifikasi yang dapat digunakan adalah metode *naive bayes* yang sering disebut sebagai *naive bayes classifier* (NBC). *Naive bayesclassifier* (NBC) merupakan salah satu metode pada teknik klasifikasi dan termasuk dalam *classifier* statistik yang dapat memprediksi probabilitas keanggotaan *class*. NBC berprinsip pada teori bayes. NBC mengasumsikan bahwa nilai atribut pada sebuah *class* adalah independen terhadap nilai pada atribut yang lain.Kelebihan NBC adalah sederhana tetapi memiliki akurasi yang tinggi [8].Ada dua tahap pada proses klasifikasi data. Tahap pertama adalah pelatihan terhadap himpunan contoh (*training example*). Sedangkan tahap kedua adalah proses klasifikasi data yang belum diketahui kategorinya.

Naive bayes atau simple bayesian *classifier* memiliki prosedur sebagai berikut[9]:

1. Setiap sample data direpresentasikan dengan n-dimensional *feature vector*, X=(X1,X2,…,Xn), dengan n dibuat dari *sample* n atribut, berturut-turut A1,A2,…,An.
2. Diandaikan terdapat m *class*, C1,C2,…,Cm. Diberikan sebuah data *sample*, X (yang tidak diketahui *class* labelnya), kemudian *classifier* akan memprediksi X ke dalam *class* yang memiliki probabilitas posterior tertinggi, Naive bayes *classifier* akan menentukan *sample* X ke dalam class Ci jika dan hanya jika
   $P(Ci \mid X) > P(Cj \mid X)$ untuk $1 \le j \le m, j \ne i$ (3)
3. *Class* Ci adalah nilai terbesar, yang disebut dengan maksimum posteri hypothesis dengan teorema bayes :
   $P(\text{Ci} \mid X) = \frac{P(X \mid Ci)\ P(Ci)}{P(X)}$ (4)
4. P(X) adalah konstan untuk semua *class*. Jika probabilitas *class* prior tidak diketahui, secara umum diasumsikan bahwa *class* adalah sama, yaitu P(C1)=P(C2)=…=P(Cm), dan selanjutnya menghitung nilai P(X|Ci) dan menghitung nilai P(X|Ci)P(Ci). Probabilitas class prior diestimasi dengan
   $P(\text{Ci}) = \frac{s_i}{s}$ (5)
   dimana s_i adalah jumlah training sample pada class Ci, dan s adalah jumlah *training sample*.
5. Apabila *dataset* terdiri dari banyak atribut, akan mengakibatkan komputasi yang rumit untuk menghitung P(X|Ci). Untuk mengurangi komputasi, naive bayes

mengasumsikan pada pembuatan *class independen*. Sehingga nilai pada atribut dikondisikan bersifat independen antara atribut yang satu dengan atribut yang lain, serta diantara atribut tidak terdapat relasi depedensi.

$P(\mathrm{X} \mid Ci) = \prod_k^n P(Xk \mid Ci)$ (6)

6. Probabilitas $P(\mathrm{X1} \mid Ci)$, $P(\mathrm{X2} \mid Ci)$,…, $P(\mathrm{Xn} \mid Ci)$, dapat diestimasi dari training sample, dimana
   a. Jika A_k adalah kategorikal, maka $P(\mathrm{x_k} \mid \mathrm{Ci}) = \frac{s_{ik}}{s_i}$ (7)
      s_{ik} adalah jumlah dari training *sample* pada *classCi* yang mempunyai nilai X_k untuk A_k dan s_i adalah jumlah *training sample* yang termasuk ke dalam *class Ci*.
   b. Jika A_k bernilai kontinyu, maka diasumsikan mempunyai sebuah gaussian distribusi
   c. $P(x_k|C_i = g(x_k, \mu C_i, \sigma C_i) = \frac{1}{\sqrt{2\pi\sigma C_i}} e^{-\frac{(x_k-\mu C_i)^2}{2\sigma^2 C_i}}$ (8)
      $g(x_k, \mu C_i, \sigma C_i)$adalah fungsi gaussian untuk atribut A_k dengan μC_i dan σC_i adalah *mean* dan standard deviasi untuk atribut A_k pada *training sample class* Ci.
7. Untuk mengklasifikasikan *sample* X yang tidak diketahui, $P(\mathrm{X} \mid Ci)$ $P(Ci)$ dievaluasi untuk setiap *class* Ci. *Sample* X ditetapkan untuk *class* Ci jika dan hanya jika
   $P(Ci \mid X) > P(Cj \mid X)$ (9)
   untuk $1 \leq j \leq m$, $j \neq i$
   Dengan kata lain, ditetapkan sebagai *class* Ci untuk $P(Ci \mid X)$ yang bernilai maksimum.

f. TextMining

Tahap awal yang dilakukan adalah melakukan preprocessing data. Tahap ini merupakan proses awal terhadap teks untuk mempersiapkan teks menjadi data yang akan diolah lebih lanjut. Twitter, sesuai dengan karakteristiknya memiliki frekuensi *slang* dan ejaan yang salah sangat tinggi maka harus diberishkan terlebih dahulu sebelum diproses lebih lanjut. Preprocessing yang dilakukan terhadap tweet seperti yang dilakukan [6] meliputi beberapa tahap antara lain casefolding, filtering,tokenasi, slang replacement dan stopword removal.

Casefolding, yakni mengubah setiap karakter huruf dalam isi dokumen menjadi huruf kecil. Tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu, peran casefolding dibutuhkan dalam mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar (biasanya huruf kacil).

Filtering, yakni mengeliminasi karakter-karakter ilegal dalam isi dokumen, antara lain menghilangkan URL Links (misalnya : http://contoh.com), Twitter Username (seperti @aku-dengan simbol @ yang mengindikasikan suatu username), Twitter special words (seperti "RT") dan emoticons. Selain itu dilakukan juga menghilangkan semua simbol dan karakter seperti %, ^,* dan lainnya.

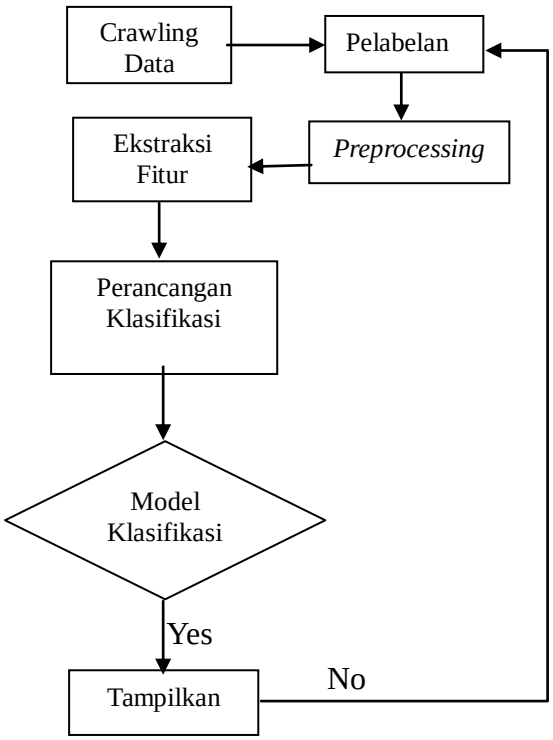
Tokenisasi, yakni melakukan pemecahan sekumpulan karakter (kalimat) menjadi kata-kata (token). Setiap token adalah objek dari suatu tipe, sehingga jumlah token akan lebih banyak daripada tipenya. Slang replacement, yakni mengganti kata-kata yang tidak sesuai dengan bahasa indonesiamenggunakan kamus lokal, misalnya : 'lemot' menjadi 'lambat'.

Stopword adalah kata umum (*common words*) yang biasanya muncul dalam jumlah besar di setiap dokumen dan dianggap tidak memiliki makna[5]. Proses penghapusan*stopword* ini dilakukan untuk setiap dokumen, apabila di dalam dokumen ditemukan kata yang termasuk kedalam daftar *stopword* maka kata tersebut dihapus, sehingga dimensi dokumen menjadi berkurang.

III.METODE PENELITIAN

Metode penelitian yang dilakukan seperti ditunjukka pada Gambar 1 adalah sebagai berikut:

1. *Crawling* Data, yaitudata twitter berisiopini masyarkat yang mengandung unsur nama calon Gubernur DKI jakarta tahun 2017 dari server Twitter memanfaatkan API *search* Twitter.
2. Pelabelan sentiment tweet menjadi netral, positif atau negative
3. *Preprocessing*, dilakukanuntuk menghasilkan data bersih. Preprocessing meliputi: casefolding, filtering,tokenasi, slang replacement dan stopword removal.
4. *Ekstraksi Fitur*, meliputi: Unigram, Negation, TF dan TF-IDF
5. Perancangan klasifikasi, dilakukan untuk mengelompokkan tweet berdasarkan kelas yang ditentukan.
6. Model Klasifikasi, digunakan untuk klasifikasi data baru, model ini didapatkan dari formula matematika (algoritma) menggunakan NBC.



Gambar 1. Alur Metodologi Penelitian

IV. Hasil danPembahasan

a. Crawling Data

Data yang dicrawling pada twitter berbahasa Indonesiadengan memanfaatkan API Search Twitter menggunakan fungsi JSON, dengan kata kunci "AHY" untuk pasangan Agus-Sylvi, "Ahok" untuk pasangan Ahok-Djarot dan "Anies" untuk pasangan Anies-Sandi. Data diambil dari data bulan November –Desember 2016 sejumlah 4401 *record*yang terdiri dari 1550 tweet untuk AHY, 1500 tweet untuk Ahok, 1351 tweet untuk Anies. Tabel 1 merupakan contoh data tweet hasil Crawling.

TABEL 1
CONTOH DATA TWEET

No	Data Tweet
1	@Raywailersrasta: Yang tua, yang remaja, yang muda semua dukung AHY #SemuaDukungAgus @AYOJakarta2017
2	@rumahkusatu: Jakarta akan dibangun Anies dengan gerakan, bukanprogram\nhttps://t.co/jrvaMsgBdH #SalamBersama #terpopuler #news @inilah
3	@TeukuDiciawi: AHY : Saya Beruntung Dapat Mpok Sylvi... #MpokSylviPelayanMasyarakat https://t.co/A7lF4WQqCS.
4	@jitunews: Nasional | Unik, Sandiaga Uno Dikira Anies BaswedanSaat Kampanye di Tanjung Priok https://t.co/1zwns6rsl9 #BeritaJitu
5	@benysalim11: Kalau Ahok Menista Agama, Mayatnya Udah Tidak utuh https://t.co/BsOqtGU5NF
6	@edo_macho:@cagubnyinyir @anto_winarno@Yuli52389173 @arifin_henry@marcel_prawira @GhazaliSil mnusia buta yg yobloz ahybs krj apa klo cuma jnji mnis

b. Pelabelan

Data tweet diberikan label berdasarkan isi tweetnya, apakah mengandung opini positif, negatif atau netral. Pelabelan opini ditunjukkan pada Table 2.

TABEL 2
CONTOH PELABELAN OPINI

No	Data Tweet	Opini
1	@Raywailersrasta: Yang tua, yang remaja, yang muda semua dukung AHY #SemuaDukungAgus @AYOJakarta2017	Positif
2	@rumahkusatu: Jakarta akan dibangun Anies dengan gerakan, bukanprogram\nhttps://t.co/jrvaMsgBdH #SalamBersama #terpopuler #news @inilah	Positif
3	@TeukuDiciawi: AHY : Saya Beruntung Dapat Mpok Sylvi... #MpokSylviPelayanMasyarakat https://t.co/A7lF4WQqCS.	Netral
4	@jitunews: Nasional | Unik, Sandiaga Uno Dikira Anies BaswedanSaat Kampanye di Tanjung Priok https://t.co/1zwns6rsl9 #BeritaJitu	Netral
5	@benysalim11: Kalau Ahok Menista Agama, Mayatnya Udah Tidak utuh https://t.co/BsOqtGU5NF	Negatif
6	@edo_macho:@cagubnyinyir @anto_winarno@Yuli52389173 @arifin_henry@marcel_prawira @GhazaliSil mnusia buta yg yobloz ahybs krj apa klo cuma jnji mnis	Negatif

c. Preprocessing

Proses ini meliputi casefolding, filtering, tokenasi, slang replacement dan stopword

removal. Hasil Preprocessing ditunjukkan pada Tabel 3.

TABEL 3.
HASIL PREPROCESSING

1. Tua muda remaja dukung AHY
2. Jakarta bangun anies gerak bukan program
3. Untung dapat sylvi
4. Sandiaga uno dikira anies baswedan kampanye tanjung priok
5. Manusia buta nyoblos AHY bias kerja apa Cuma janji manis
6. Ahok nista agama mayat sudah tidak utuh

d. Perancangan Klasifikasi dengan Algoritma Naïve Bayes Classifier

Setelah mengalami preprocessing dan seleksi fitur, data tweet tersebut diklasifikasi menggunakan algoritma naïve bayes Classifier, klasifikasi berdasarkan sentiment atau opini positif, negative atau netral. Hasil klasifikasi sentiment ditunjukkan pada Tabel5, sedangkan untuk pelabelan manual ditunjukkan pada Table 4.

TABEL 4
HASIL PELABELAN MANUAL

Query	Pelabelan Manual			Total
	Positif	Negatif	Netral	
AHY	547	498	505	1550
Ahok	1237	263	0	1500
Anies	981	302	68	1351

TABEL 5
HASIL KLASIFIKASI DENGAN NAÏVE BAYES CLASSIFIER

Query	Hasil Klasifikasi			Total
	Positif	Negatif	Netral	
AHY	550	493	512	1550
Ahok	1231	258	11	1500
Anies	1000	300	51	1351

V. KESIMPULAN

Berdasarkan percobaan-percobaan yang dilakukan dapat disimpulkan sebagi berikut:

1. Crawling menggunakan JSON relatif lebih mudah karena sudah tersedia pada aplikasi twitternya
2. Hasil nilai sentiment positif antara klasifikasi dan pelabelan manual memiliki selisih yang cukup siginifikan. yaitu pelabelan AHY positif 547, Klasifikasi AHY Positif 550. Untuk pelabelan Ahok sentimen positif 1237 dan hasil klasifikasi positif 1231. Sedangkan pelabelan Anies sentimen positif 981 dan hasil klasifikasi positif 1000.
3. Hasil nilai sentiment Negatif antara klasifikasi dan pelabelan manual tidak terlalu siginifikan. yaitu pelabelan AHY negative 498, Klasifikasi AHY negative 493. Untuk pelabelan Ahok sentimen negative 263 dan hasil klasifikasi positif 258. Sedangkan pelabelan Anies sentimen negatif 302 dan hasil klasifikasi positif 300.
4. Hasil nilai sentiment netral antara klasifikasi dan pelabelan manual memiliki selisih yang cukup siginifikan. yaitu pelabelan AHY netral 505, Klasifikasi AHY netral 512. Untuk pelabelan Ahok sentimen netral 0 dan hasil klasifikasi netral 11. Sedangkan pelabelan Anies sentimen netral 68 dan hasil klasifikasi netral 51
5. Data yang sudah mengalami preprocessing dan diekstrak akan lebih mudah pengklasifikasiannya/pelabelan, dibandingkan data yang belum dipreprocessing dan juga diekstrak

VI. REFERENSI

[1] M. Yarrow, J. Clausen, and P. Robbins, "The social meaning of mental illness," *J. Soc. Issues*, pp. 443–454, 2010.

[2] M. Huberty, "Can we vote with our tweet? On the perennial difficulty of election forecasting with social media," *Int. J. Forecast.*, vol. 31, no. 3, pp. 992–1007, 2015.

[3] B. Liu, *Handbook of Natural Language Processing*, 2ndEdition ed. Taylor and Francis Group, LLC Chapman, 2010.

[4] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," in *Foundations and Trends® in Information Retrieval*, vol. 1, no. 2, 2008, pp. 91–231.

[5] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval Introduction*, vol. 35, no. 2. 2008.

[6] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Processing*, vol. 150, no. 12, pp. 1–6, 2009.

[7] E. Blanco and D. Moldovan, "Some Issues on Detecting Negation from Text," *Twenty-Fourth Int. FLAIRS Conf.*, pp. 228–233, 2011.

[8] I. Rish, "An empirical study of the naive Bayes classifier," *Empir. methods Artif. Intell. Work. IJCAI*, vol. 22230, no. January 2001, pp. 41–46, 2001.
[9] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 54, no. Second Edition. 2006.