ANALISA SENTIMEN TWITTER PADA PILPRES 2019 MENGGUNAKAN ALGORITMA NAIVE BAYES

Lia Durrotul Mahbubah[1], Eri Zuliarso[2]
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Stikubank
E-mail: [1]liadm7106@gmail.com, [2]eri299@edu.unisbank.ac.id

Abstrak - Indonesia adalah negara yang menganut sistem demokrasi. Hal ini ditandai dengan diadakannya suatu pemilihan umum terhadap presiden dan wakil presiden. Salah satu kandidat calon presiden indonesia adalah Prabowo Subianto. Sebagai seorang tokoh politik pasti mendapat berbagai komentar atau opini dari masyarakat. Di zaman sekarang banyak masyarakat yang mengekspresikan opininya pada media sosial twitter. Pada penilitian ini, akan mengambil tweets dari twitter dengan kata kunci pencarian #pilpres2019 dan #prabowo untuk diolah dan mengklasifikasikan teks dengan menggunakan metode analisis sentimen. Untuk proses klasifikasi teks dibagi menjadi dua kelas yaitu kelas sentimen positif dan kelas sentimen negatif. Data yang digunakan berjumlah 300 tweets yang terdiri dari 240 data latih dan 60 data uji. Data untuk pelatihan sudah diketahui sentimennya sedangkan data untuk pengujian belum diketahui nilai sentimennya. Dari 240 data latih terdiri dari 134 sentimen negatif dan 106 sentimen positif. Pada studi ini menunjukkan bahwa klasifikasi data tweets menggunakan algoritma naive bayes classifier memberikan akurasi sebesar 73%. Precision kelas negatif sebesar 78% dan precision kelas positif sebesar 66%.
Kata kunci – *Naive Bayes*, Analisis Sentimen, Klasifikasi, Twitter, Pilpres 2019, Prabowo Subianto

1. PENDAHULUAN

Indonesia adalah salah satu negara yang menganut sistem demokrasi. Hal ini ditandai dengan diadakannya suatu pemilihan umum terhadap presiden dan wakil presiden. Pemilihan umum pada suatu negara yang menganut demokrasi biasanya diselenggarakan secara periodik. Pada tahun 2019 akan dilaksanakan pemilihan umum presiden dan wakil presiden. Seorang tokoh politik yang ingin maju sebagai calon presiden tentu akan melihat atau mempertimbangkan popularitas mereka berdasarkan opini dari masyarakat. Dahulu masyarakat mengungkapkan opini, kritik, dan sarannya melalui media cetak yang tidak semua orang mempunyai kemampuan menulis dan kesempatan menerbitkan tulisannya. Namun, perkembangan teknologi komunikasi saat ini telah merubah kecenderungan kebiasaan masyarakat dalam mengekspresikan opininya pada media sosial. Salah satu media sosial yang populer di kalangan pengguna internet saat ini adalah Twitter.

Twitter merupakan media sosial yang dibuat oleh Jack Dorsey pada tahun 2006. Pada tahun 2019 Berdasarkan *press-release* Twitter ada 500 juta *tweet* atau kicauan oleh pengguna twitter per harinya[1]. Sebanyak 500 juta tweet digunakan untuk mem-post hal tentang diri pengguna dan berbagi informasi, isi *tweet* juga dapat mengekspresikan perasaan[2]. Opini melalui *tweet* inilah yang dapat dimanfaatkan untuk melihat bagaimana sentimen yang dimunculkan salah satunya adalah mengenai opini seseorang terhadap tokoh politik yang akan maju sebagai calon presiden Indonesia tahun 2019.

Penentukan polaritas positif atau negatifnya suatu opini dapat dilakukan secara manual, tetapi seiring bertambahnya sumber opini menjadi semakin banyak tentunya waktu dan usaha yang dibutuhkan untuk mengklasifikasikan polaritas opini tersebut akan semakin banyak terpakai. Oleh karena itu, diajukan penerapan metode pembelajaran mesin untuk mengklasifikasi polaritas opini dari sumber data yang sangat banyak tersebut. Untuk melakukan hal itu, bisa menggunakan salah satu fungsi dari text mining, dalam hal ini adalah klasifikasi dokumen.

Ada beragam teknik klasifikasi dokumen, di antaranya adalah *Naïve Bayes classifier*, *Decision Trees*, dan *Support Vector Machines*. Salah satu metode yang paling populer digunakan dalam pengklasifikasian dokumen sekarang ini adalah metode *Naïve Bayes classifier*[3]. Metode *Naïve Bayes classifier* mempunyai kecepatan dan akurasi yang tinggi ketika diaplikasikan dalam basis data yang besar dan data yang beragam[4]. *Metode Naïve Bayes Classifier* memiliki beberapa kelebihan antara lain, sederhana, cepat dan berakurasi tinggi[5].

Sehingga dalam menyelesaikan permasalahan tersebut dilakukan "Analisis Sentiment Pada Media Sosial Twitter Menggunakan Algoritma *Naive Bayes*", terhadap tokoh politik yang maju dalam pemilihan umum presiden dan wakil presiden Indonesia tahun 2019.

2. METODE PENELITIAN

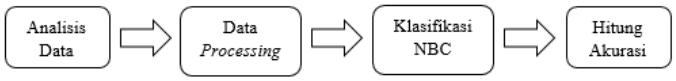
2.1 Metode Pengumpulan Data

Beberapa metode untuk memperoleh data atau informasi dalam menyelesaikan permasalahan. Metode yang dilakukan tersebut antara lain :

1. Observasi
   Metode observasi yang dilakukan adalah dengan pengamatan langsung pada obyek untuk mendapatkan gambaran yang lengkap mengenai data yang akan diproses
2. Data yang diperoleh berupa data primer yaitu data yang diperoleh secara langsung dari sumber atau objek penelitian. Data tersebut merupakan opini masyarakat terhadap calon presiden Indonesia 2019 di twitter dengan kata kunci pencarian #pilpres2019 dan #prabowo.

2.2 Metode Analisis Data

Metode yang digunakan untuk menganalisis data *tweet* untuk mengetahui kelas sentimennya adalah sebagai berikut :



Gambar 1. Alur Sistem Analisis Sentimen

Berikut adalah penjelasan mengenai alur sistem analisis sentimen pada gambar 1 :

1. Analisis Data
   Setelah memperoleh data dari twitter sebanyak 300 *tweet,* selanjutnya data tersebut dibagi menjadi dua yaitu data latih sebanyak 240 *tweet* dan data uji sebanyak 60 *tweet.* dari 240 data latih, kemudian ditentukan nilai sentimennya secara manual yaitu dengan 106 *tweet* sentimen positif dan 134 *tweet* sentimen negatif.
2. Data Processing
   Data processing atau text processing berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur. Secara umum proses yang dilakukan dalam tahapan preprocessing adalah *case folding*, normalisasi, *tokenizing*, dan *stopword removal.*
3. Klasifikasi *Naive Bayes Classifier*
   Naïve Bayes Classifier (NBC) merupakan sebuah metode klasifikasi yang berakar pada *teorema Bayes*. Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing variabel. Tahap ini adalah proses pengklasifikasian menggunakan algoritma *naive bayes classifier* untuk menentukan data tersebut termasuk dalam opini positif atau opini negatif.
4. Hitung Akurasi
   Data yang digunakan untuk menghitung akurasi adalah data uji yang sudah diketahui sentimennya. Untuk menghitung tingkat akurasi dari klasifikasi sistem menggunakan *confusion matrix*. Data hasil klasifikasi akan dihitung berapa banyak data yang benar dan berapa banyak data yang salah.

2.3 Metode Klasifikasi *Naive Bayes Classifier*

Secara umum proses ini dibagi menjadi beberapa tahap. Antara lain sebagai berikut :

1. Data uji yang belum diklasifikasi akan melalui tahap *preprocessing text*. Tahapan *preprocessing text* terdiri dari *case folding*, normalisasi, *tokenizing*, dan *stopword removal*.
2. Setelah dilakukan *preprocessing* maka data *tweet* selanjutnya akan dilakukan penghitungan frekuensi *term*.
3. Kemudian menghitung nilai Vmap untuk tiap kelas dengan persamaan rumus :

$$V_{map} = \underset{V_j \in V}{argmax} \prod_{i=1}^{n} P(xi|V_j) P(V_j) \quad (1)$$

Keterangan :

V_j Kategori tweet j =1, 2,…n. Dimana dalam penelitian ini
j1 = kategori tweet sentimen positif,
j2 = kategori tweet sentiment negatif

$P(xi|V_j)$ Probabilitas xi pada kategori V_j

$P(V_j)$ Probabilitas dari V_j

Kelas suatu *tweet* ditentukan berdasarkan nilai Vmap terbesar.

4. Hitung akurasi
   Untuk menghitung tingkat akurasi dari klasifikasi sistem menggunakan *confusion matrix*. Data hasil klasifikasi akan dihitung berapa banyak data yang benar dan berapa banyak data yang salah. Untuk menghitung tingkat akurasi maka menggunakan persamaan rumus :

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

Keterangan :
TP : True Positif TN : True Negatif
FP : False Positif FN : False Negatif

3. HASIL DAN PEMBAHASAN

Data *tweet* diperoleh dari media sosial Twitter dengan kata kunci pencarian #pilpres2019 dan #prabowo. Data yang diunduh dari Twitter berupa *tweet* bahasa Indonesia dengan jumlah data 300 *tweet.* kemudian data akan disimpan ke dalam database dengan format *xlsx.* Dari 300 data *tweet* tersebut di bagi menjadi dua yaitu data latih dan data uji. Pembagian data latih dan data uji dapat dilihat pada Tabel 1.

Tabel 1. Pembagian Data

Jenis Sentimen Tweet	Negatif	Positif
Latih	134	106
Uji	35	25

3.1 Pembacaan Data

Setelah menentukan pembagian data latih dan data uji, selanjutnya pembacaan data ke dalam RStudio. Gambar 1 adalah partisi data latih, gambar 2 adalah partisi dari data uji.

	class	text	username	tgl
122	neg	V_Stone_Kardol @RamliRizal Ini Bu Sri Mulyani dpt peng...	@V_Stone_Kardol	2019-01-29
123	pos	Pertemuan Jokowi dan Prabowo untuk kepentingan Ba...	@Rabiatul173	2019-07-15
124	pos	Kebersamaan kemaren #Pilpres2019 @jokowi bareng @...	@Kurniaw35770172	2019-07-09
125	neg	kompascom klo sudah tau itu Harus nya #AlimUlama me...	@kompascom	2019-01-29
126	neg	Ditahan karena 1. Publik figur 2. Suaranya menndukung...	@kla_110	2019-01-28
127	neg	Dan bila kemudian hasil Quick Count #Pilpres2019 tidak...	@Rabbanildeas	2019-04-20
128	neg	Konflik internal Gerindra semakin rumit pasca #Pilpres20...	@OtakKampret_	2019-07-07
129	pos	#RekonsiliasiJokowiPrabowo Menyejukkan Politik Tanah...	@dinasukmara	2019-07-06
130	neg	FPI Udah Mundur Mendukung Prabowo Dikarenakan Pr...	@matamilenial	2019-01-29
131	pos	Bedah Visi Capres di UGM Dapat Pujian #sandiagauno ...	@2019pastiPAS	2019-01-31
132	neg	MohdNur_Thahir @bertusbeben @Dahnilanzar Ratna S...	@Tarikh_Tambang	2019-01-29
133	pos	Ga heran beliau menang #Pilpres2019 , menang #Putus...	@cutsarah	2019-07-05

Showing 121 to 133 of 240 entries

Gambar 2. Data Latih

	class	text	username	tgl
26	pos	Justru sebaliknya, @prabowo bertemu saat semua prose...	@ConsultantKrci	2019-07-14
27	pos	Prabowo: Pertemuan Saya Dengan Presiden Jokowi Unt...	@adelia_dara	2019-07-14
28	pos	Kami Jamin Keamanan Kotak Suara Pemilu 2019 . . #TNIP...	@JayapuraKota	2019-04-22
29	neg	Bukan kalah tapi di kalahkan...#PrabowoJanganHadiriP...	@Kurniaw35770172	2019-07-06
30	neg	baguslah, tp jgn bentak2 reporter. hostnya ya pakkk :D ...	@rahabganendra	2019-01-30
31	neg	ucapan dari #Prabowo soal Menteri Keuangan sebagai ...	@Kitabicar4	2019-01-30
32	pos	Jadi siapa yang menang #Pilpres2019 ? Kunjungan pakd...	@Elina_Vay	2019-04-20
33	neg	Penyebar hoax harusnya diproses hukum ... klo gak ora...	@hanifah932	2019-04-27
34	pos	Senyum tipis menjadi respons Menkeu Sri Mulyani terkai...	@detikfinance	2019-01-29
35	neg	Beritasatu Ditolak FMB, Festival Cap Go Meh Tetap Didu...	@Beritasatu	2019-01-29
36	pos	Silaturahmi Jokowi-Prabowo di Bulan Juli Bikin Adem Su...	@HaluanKita	2019-07-06
37	pos	Pemikiran asli pk @prabowo #2019GantiPresiden #2019...	@Fatsp18	2019-01-28

Showing 26 to 37 of 60 entries

Gambar 3. Data Uji

3.2 *Text Processing*

Langkah selanjutnya yaitu *text processing*, tindakan yang dilakukan pada tahap ini adalah tokenisasi, mengubah semua huruf menjadi huruf kecil, menghapus tanda baca, menghapus angka dan menghapus kata – kata (*stopword removal*) yang tidak memiliki makna sehingga tidak berpengaruh terhadap proses analisis sentimen. *Script* untuk pemrosesan teks dapat dilihat pada Gambar 3.

```
corpus.clean <- corpus %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(removePunctuation) %>%
  tm_map(removeNumbers) %>%
  tm_map(removeWords, stopwords(kind="en")) %>%
  tm_map(stripWhitespace)
```

Gambar 4. *Script* Pemrosesan Teks

3.3 Pelatihan Model *Naive Bayes*

Untuk melatih model, penelitian ini menggunakan fungsi *naive bayes* dari paket 'e1071'. Karena *Naive Bayes* mengevaluasi probabilitas, maka memerlukan beberapa cara untuk menetapkan probabilitas tidak nol pada kata-kata yang tidak muncul dalam sampel. Dengan menggunakan *Laplace* 1 *smoothing* untuk tujuan ini.Gambar 4 merupakan *script* model *naive bayes classifier* sedangkan hasil klasifikasi data uji dapat dilihat pada Gambar 5.

```
trainNB <- apply(dtm.train.nb, 2, convert_count)
testNB <- apply(dtm.test.nb, 2, convert_count)

system.time( classifier <- naiveBayes(trainNB, df.train$class, laplace = 1) )

predict(classifier, newdata=testNB)
system.time( pred <- predict(classifier, newdata=testNB) )

table("Predictions"= pred,  "Actual" = df.test$class )
```

Gambar 5. Model *Naive Bayes Classifier*

```
> predict(classifier, newdata=testNB)
 [1] neg neg pos neg pos pos pos neg neg neg neg neg neg neg neg
[16] pos pos neg pos pos pos neg pos neg neg pos pos neg neg neg
[31] neg pos neg pos pos pos pos pos pos neg neg neg neg neg neg
[46] neg neg pos neg pos pos pos pos pos neg neg neg pos pos neg
Levels: neg pos
```

Gambar 6. Hasil Klasifikasi

3.4 Hitung Akurasi

Dari pengujian 60 data tweet algoritma *naive bayes classifier* pada Gambar 5 memberikan hasil klasifikasi tweet ke dalam sentimen neg (negatif) sebanyak 33 dokumen dan sentimen pos (positif) sebanyak 27 dokumen. Dari hasil prediksi tersebut akan dihitung tingkat akurasi algoritma *naive bayes classifier* dalam melakukan klasifikasi *tweet.* Untuk menghitung tingkat akurasi dari klasifikasi sistem menggunakan *confusion matrix.* Data hasil klasifikasi algoritma *naive bayes* akan dicocokan dengan sentimen yang telah diketahui sebelumnya. Hasil dari *Confusion Matrix* dapat dilihat pada Gambar 6.

```
> table("Predictions"= pred,  "Actual" = df.test$class )
           Actual
Predictions neg pos
        neg  26   7
        pos   9  18
```

Gambar 7. Hasil *Confusion Matrix*

Dari gambar 6 hasil pencocokan data sentimen klasifikasi algoritma *naive bayes* dengan sentimen sebenarnya menghasilkan :

1. *True* Positif 18 dokumen
2. *True* Negatif 26 dokumen
3. *False* Positif 9 dokumen
4. *False* Negatif 7 dokumen

```
               Accuracy : 0.7333
                 95% CI : (0.6034, 0.8393)
    No Information Rate : 0.5833
    P-Value [Acc > NIR] : 0.0116

                  Kappa : 0.4576

 Mcnemar's Test P-Value : 0.8026

            Sensitivity : 0.7429
            Specificity : 0.7200
         Pos Pred Value : 0.7879
         Neg Pred Value : 0.6667
             Prevalence : 0.5833
         Detection Rate : 0.4333
   Detection Prevalence : 0.5500
      Balanced Accuracy : 0.7314

       'Positive' Class : neg
```

Gambar 8. Akurasi Klasifikasi *Naive Bayes*

Pada Gambar 8 akurasi dari algoritma *naive bayes* dalam melakukan klasifikasi *tweet* dalam penelitian ini sebesar 0.73 = 73%. Akurasi dari penelitian ini menunjukkan bahwa meskipun banyak asumsi yang disederhanakan, algoritma Naive Bayes cukup baik dalam memprediksi kelas sentimen yang benar.

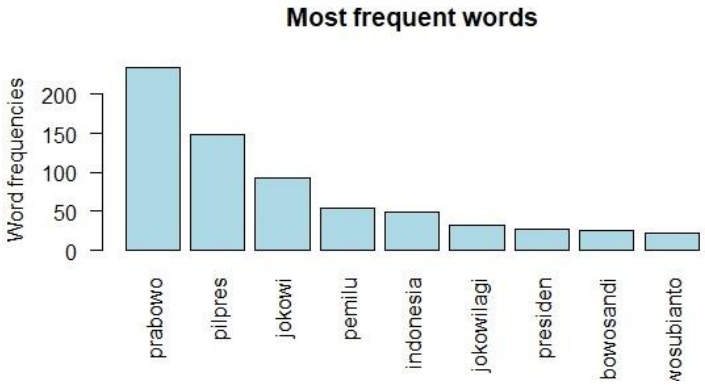
3.5 Visualisasi

Tahap ini adalah tahapan yang menampilkan data kata – kata (*term*) yang paling sering muncul. Visualisasi yang digunakan adalah wordcloud dan diagram batang.

	word	freq
prabowo	prabowo	234
pilpres	pilpres	148
jokowi	jokowi	93
pemilu	pemilu	55
indonesia	indonesia	49
jokowilagi	jokowilagi	32
presiden	presiden	27
prabowosandi	prabowosandi	26
prabowosubianto	prabowosubianto	23

Gambar 9. Data Visualisasi

Gambar 10. Visualisasi Wordcloud



Gambar 11. Visualisasi Diagram Batang

Dalam gambar 9 sampai gambar 11 dapat dilihat 9 daftar kata yang paling sering muncul di dalam data *tweet*. Daftar 9 kata dengan kemunculan yang paling sering yaitu prabowo, pilpres, jokowi, pemilu, indonesia, jokowilagi, presiden, prabowosandi, dan prabowosubianto.

3.6 Hasil Penelitian

Dari hasil *confusion matrix* tersebut, maka didapatkan akurasi sebesar :

$$\frac{TP+TN}{TP+FP+TN+FN} = \frac{18+26}{18+9+26+7} = \frac{45}{60} = 0.73$$

Hasil akurasi algoritma *naive bayes classfier* sebesar 0,73 atau 73% dalam mengklasifikasikan *tweet* ke dalam sentimen kelas negatif dan positif. Dengan hasil *confusion matrix* tersebut, maka didapat hasil *precision* untuk masing-masing kelas dan *rate error* sistem sebagai berikut :

a. *Precision* Kelas Negatif

$$\frac{TN}{TN+FN} = \frac{26}{26+7} = \frac{26}{33} = 0.78$$

Tingkat ketepatan kelas sentimen negatif sebesar 0.78 atau 78%, sehingga tingkat kesalahan klasifikasi kelas yaitu sebesar 22%.

b. *Precision* Kelas Positif

$$\frac{TP}{TP+FP} = \frac{18}{18+9} = \frac{18}{27} = 0.66$$

Tingkat ketepatan kelas sentimen positif sebesar 0.66 atau 66%, sehingga tingkat kesalahan klasifikasi kelas yaitu sebesar 34%.

a. *Error Rate* Sistem

$$\frac{FP+FN}{TP+TN+FN+FP} = \frac{9+7}{18+26+7+9} = \frac{16}{60} = 0.26$$

Berdasarkan hasil dari *presicion* kelas negatif dan *precision* kelas positif, tingkat kesalahan mengklasifikasikan kelas sentimen dipengaruhi oleh tingkat akurasi dari jumlah data positif yang benar diklasifikasi sebagai data positif dan jumlah data negatif yang benar diklasifikasi sebagai data negatif. Sedangkan hasil dari *error rate* sistem, kesalahan hasil prediksi klasifikasi yang terdapat dalam tabel *confusion matrix* dipengaruhi oleh tidak seimbangnya jumlah data antara kelas sentimen negatif dengan kelas sentimen positif yang ada dalam data latih mengakibatkan sedikitnya referensi untuk kata-kata positif yang akan berpengaruh pada tahap klasifikasi untuk data uji mengalami beberapa kesalahan prediksi.

4. KESIMPULAN

Dari hasil penelitian, dapat disimpulkan beberapa hal sebagai berikut :

1. Metode *Naive Bayes Classifier* dalam melakukan klasifikasi *tweet* sentimen negatif dan positif dengan 240 data latih dan 60 data uji mendapat hasil akurasi sebesar 73%.
2. Dengan hasil akurasi yang cukup tinggi yaitu 73% maka metode *Naive Bayes Classifier* cukup efektif dalam melakukan klasifikasi *tweet* dengan sentimen negatif dan positif secara otomatis.
3. Analisis Sentimen terbukti dapat digunakan untuk mengetahui sentimen masyarakat khususnya pengguna twitter terhadap Calon Presiden Indonesia 2019, sehingga membantu masyarakat awam untuk mengetahui sentimen masyarakat lainya terhadap Calon Presiden Indonesia 2019.

SARAN

Dari hasil penelitian yang telah dilakukan, sistem yang dibangun masih memiliki kekurangan baik dari segi fungsionalitas maupun data yang dimiliki. Oleh karena itu agar didapat sistem yang lebih handal dan akurat perlu dilakukan pengembangan dan penyempurnaan lebih lanjut. Adapun saran agar sistem dapat berfungsi dengan lebih baik lagi yaitu :

1. Melakukan perbandingan dengan algoritma klasifikasi lain, untuk mengetahui algoritma yang memiliki akurasi terbaik.
2. Melakukan penambahan jumlah data latih dan data uji untuk mendapatkan hasil yang lebih baik saat klasifikasi *tweet.*
3. Menggunakan algoritma selain *Naive Bayes Classifier* seperti *Support Vector Machine* atau algoritma klasifikasi teks yang lainnya.

DAFTAR PUSTAKA

[1] Lin, Ying, 2019, 10 Twitter Statistics Every Marketer Should Know in 2019, *https://www.oberlo.com/blog/twitter-statistics*, 30 Juli 2019. [Online]. [Diakses 1 Agustus 2019].

[2] Bolen, Johan., Mao, Huina., Zeng, Xiaojung. , 2011,Twitter mood predicts the stock market, *Journal of Computational Science*, 2, pp.1–8.

[3] Natalius, Samuel., 2010,Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen, *Makalah II2092 Probabilitas dan Statistik*.

[4] Larose, D. T. , 2006, Naïve Bayes Estimation and Bayesian Networks, *in Data Mining Methods and Models*, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471756482.ch5.

[5] McCue, Rita. , 2009, A Comparison of the Accuracy of Support Vector Machine and Naıve Bayes Algorithms In Spam Classification, *University of California at Santa Cruz.*