Analysis of Community Sentiments Regarding Plans to Relocate National Capital Using the Naïve Bayes Method

Analisa Sentimen Masyarakat Tentang Rencana Pemindahan Ibukota Negara Dengan Metode Naïve Bayes

Tomi Eko Hidayat *, Mochamad Alfan Rosid, Ika Ratna Indra Astutik

Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

*Coresponding author.
E-mail addresses: tomiehidayat@umsida.ac.id, alfanrosid@umsida.ac.id, ikaratna@umsida.ac.id

Abstract. *The development of social media at this time can not be stopped. Social media has been used by most Indonesian people. Moving the capital to a hot issue discussed on social media. The government plans to move the capital to the island of Borneo. Of course, news about the news was much responded by netizens on social media. In this study the author tries to dig deeper into community sentiments about the plan to move the capital by utilizing the comments page on Twitter Media. The method used in this research is Naïve Bayes Classifier, a classic method that has a pretty good accuracy. Naive Bayes Classifier is a probabilistic classification based on the Bayes theorem, taking into account naïv independence assumptions. In addition to using the naïve bayes method, in this study the researchers also used word weighting. The weighting word used is TF-IDF, which is a combination of term frequency and inverse document frequency. By using 3 testing methods, namely Confusion matrix, Precission and Recall, and K-Fold Cross Validation. The results obtained in this study are 3 document classifications, namely Positive, Negative and Neutral. Testing is done by dividing the document into 2 subsets, namely training data and test data and the resulting accuracy of 64.6%.*
Keywords: Text Mining; Naïve Bayes Classifier; TF-IDF; Sentiment Analysis

Abstrak. *Perkembangan media sosial saat ini memang sudah tidak bisa terbendung. Media sosial telah digunakan oleh sebagian besar masyarakat Indonesia. Pemindahan ibukota menjadi isu yang hangat dibicarakan di media sosial. Pemerintah berencana memindahkan ibukota ke pulau Kalimantan. Tentunya kabar tentang pemberitaan tersebut banyak ditanggapi oleh netizen di media social. Pada penelitian ini penulis mencoba menggali lebih dalam sentimen masyarakat tentang rencana pemindahan ibukota dengan memanfaatkan laman komentar pada Media Twitter. Metode yang digunakan dalam penelitian ini yaitu dengan Naïve Bayes Classifier, metode klasik yang mempunyai akurasi yang lumayan bagus. Naive Bayes Classifier merupakan pengklasifikasian probabilistik berdasarkan teorema bayes, mempertimbangkan asumsi kemandirian naïv, .Selain dengan menggunakan metode naïve bayes, pada penelitian ini peneliti juga menggunakan pembobotan kata. Pembobobtan kata yang dipakai adalah TF-IDF yaitu penggabungan antara term frequency serta inverse document frequency. Dengan menggunakan 3 metode pengujian yaitu Confussion matrix, Precission and Recall, serta K-Fold Cross Validation. Hasil yang didapat pada penelitian ini yaitu 3 klasifikasi dokumen yaitu Positif, Negatif dan Netral. Pengujian dilakukan dengan membagi dokumen menjadi 2 subset, yaitu data latih serta data uji dan dihasilkan akurasi sebesar 64,6%.*
Kata kunci- *Teks Mining; Naïve Bayes Classifier; TF-IDF; Analisa Sentimen, Ibu Kota*

PENDAHULUAN

Pemindahan ibukota menjadi isu yang hangat dibicarakan di media sosial. Pemerintah berencana memindahkan ibukota ke pulau Kalimantan. Dikutip dari laman berita liputan 6 pada tanggal September 26, 2019, presiden Jokowi membenarkan tentang pemindahan ibukota tersebut. Namun untuk kota yang akan dijadikan sebagai ibukota masih dilakukan analisa lebih dalam oleh pemerintah. Ibukota Negara sendiri merupakan kota yang menjadi tempat pusat kedudukan suatu Negara baik administratif, legislatif, eksekutif, dan yudikatif suatu Negara. Tentunya kabar tentang pemberitaan tersebut banyak ditanggapi oleh netizen di media sosial[1].

Perkembangan media sosial saat ini memang sudah tidak bisa terbendung[2][3]. Media sosial telah digunakan oleh sebagian besar masyarakat Indonesia. Dikutip dari laman resmi kominfo pada September 26, 2019 bahwa pengguna internet di Inodenesia sebanyak 63 Juta pengguna dan 95% pengguna tersebut menggunakan internet sebagai media sosial. Menurut data dari PT Bakrie Telecom, twitter memiliki 19,5 Juta Pengguna di Indonesia.

Pada penelitian ini penulis mencoba menggali lebih dalam sentimen masyarakat tentang rencana pemindahan ibukota. Menurut Kamus Besar Bahasa Indonesia, sentimen merupakan pendapat yang didasarkan atas perasaan yanqg berlebihan terhadap sesuatu. Sentimen disini dilontarkan kedalam bentuk tulisan di media sosial dan merupakan komentar publik terhadap kondisi tertentu.

Komentar tersebut akan diolah untuk menjadi informasi dengan menggunakan suatu tools dan menghasilkan data mining berupa teks. Teks mining mempunyai tujuan untuk menggali data dan informasi dari beberapa dokumen. Sumber data yang digunakan dalam teks mining berasal dari kalimat atau sekumpulan teks yang memiliki format tidak terstruktur ataupun semi terstruktur[4].

Untuk mengolah data ini penulis menggunakan metode Naïve Bayes untuk menentukan sentimen positif , negative, dan netral terkait masalah diatas. Naïve bayes juga sudah digunakan oleh Agnes Rossi Trisna Lestari untuk menganalisa Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji[5]. Dari penelitian tersebut didapatkan tingkat akurasi sistem pada pembobotan tekstual sebesar 68,52%, pada pembobotan nontesktual 75,93%, dan pada penggabungan kedua pembobotan 74,81%, dengan kesimpulan penggabungan kedua pembobotan dapat menambah akurasi sistem. Selain itu penelitian oleh Didik Garbian Nugroho juga menggunakan metode naïve bayes untuk menganalisa sentimen publik terhadap ojek online. Diperoleh kesimpulan dari penelitian tersebut yakni tingkat akurasi mencapai 80% berdasarkan 800 data tweet yang terdiri dari 300 data latih dan 500 data uji[6].

Naïve Bayes Classifier memang banyak digunakan untuk melakukan penelitian terhadap analisa sentimen publik di suatu website maupun media sosial. Naïve bayes classifier merupakan teknik pembelajaran mesin yang berbasis probabilistik[7][8]. Naïve bayes merupakan metode yang cukup sederhana untuk klasifikasi terhadap teks namun memiliki akurasi dan performasi yang cukup tinggi. Analisa sentimen yang dibangun dengan menggunakan metode naïve bayes memiliki akurasi 83%. Perbandingan metode Naïve Bayes, KNN, dan gabungan dari K-Means dan LVQ dalam mengklasifikasi kategori buku berbahasa Indonesia dengan jumlah data 200 buku yang terdiri dari 150 buku sebagai data latih dan 50 buku sebagai data uji. Hasil dari penelitian tersebut didapatkan tingkat akurasi metode KNN sebesar 96%, Naïve Bayes 98%, dan K-Means kombinasi LVQ 92,2%.

Perbandingan beberapa metode seperti metode Naïve Bayes, K-nearest Neighbor, dan gabungan K-means dan LVQ dalam mengklasifikasikan kategori buku berbahasa Indonesia dengan data yang digunakan berjumlah 200 buku dengan variabel berupa judul buku dan synopsis buku dan dibagi menjadi 150 buku digunakan sebagai data latih, sedangkan 50 buku digunakan sebagai data uji. Dari hasil penelitian yang dilakukan, metode KNN memperoleh akurasi sebesar 96%, kemudian Naïve Bayes sebesar 98%, lalu kombinasi K-Means dan LVQ menghasilkan akurasi sebesar 92,2%. Naïve bayes mendapatkan hasil akurasi tertinggi [2]. Dari penelitian yang dilakukan oleh Ni Wayan Sumartini Saraswati tentang sentimen analisis dihasilkan bahwa metode SVM dan NBC memiliki tingkat akurasi yang sama baiknya. SVM unggul pada klasifikasi opini positif sedangkan Naïve Bayes Classifier lebih unggul pada klasifikasi opini negatif, data yang digunakan yaitu opini Bahasa Indonesia pada twitter[4].

Berdasarkan latar belakang ini penulis melakukan penelitian terhadap sentiment masyarakat di media sosial twitter tentang pemindahan ibukota negara dengan menggunakan metode Naïve Bayes Classifier. Naïve bayes classifier dipilih karena memiliki metode yang tidak rumit namun bisa menghasilkan akurasi yang baik. Hasil yang akan diperoleh dari penelitian ini berupa klasifikasi sentimen masyarakat yang terbagi menjadi 3 klasifikasi yaitu positif, netral, dan negatif.

METODE PENELITIAN

A. Pengumpulan Data

Metode pengumpulan data yang digunakan adalah dengan 2 cara, yaitu studi literasi yang berasal dari buku serta jurnal tentang Analisa sentimen, Pembobotan kata, serta teks mining. Yang kedua melakukan proses crawling data twitter dengan menggunakan RStudio dengan kata kunci ibukota baru dan ibukota pindah dengan mention twitter kepada akun @jokowi, @tribunnews, @kompas dan @pak_jk. Data tersebut nantinya akan dibagi menjadi 2 bagian, yaitu data latih dan data uji, sebelum itu, data tersebut dilakukan tahap pre-processing atau pembersihan data.

B. Tahap *Preprocessing*

Tahap preprocessing disini adalah pembersihan data, fungsi dari tahap ini adalah agar akurasi yang didapat menjadi baik. Tahap preprocessing terdiri dari 4 bagian, yaitu[9][10] :

a. Tahap Case Folding : Tahap ini merupakan proses mengubah seluruh kalimat menjadi huruf kecil
b. Tahap Tokenizing : Tahap ini adalah tahap pemenggalan kalimat menjadi string/kata.
c. Tahap Stemming : Tahap ini merupakan merubah kata menjadi kata dasarnya.
d. Tahap Tagging : Tahap ini hanya dilakukan pada dokumen yang memiliki kata lampau, seperti Bahasa inggris, untuk dokumen berbahasa Indonesia tidak dilakukan tahap ini.

C. Pembobolan Kata

Pembobotan kata adalah pemberian nilai pada tiap kata berdasarkan indeks. Pembobotan kata yang digunakan pada penelitian ini adalah dengan TF IDF. TF IDF merupakan hasil perkalian dari Term Frequency atau jumlah kemunculan kata pada tiap dokumen serta Inverse Document Frequency atau kemunculan sebuah term dalam dokumen yang paling sedikit. Rumus dari TF IDF yaitu :

$$W_{t,d} = W_{tft,d} \times idf_t \qquad (1)$$

$W_{tft,d}$: Nilai term frequency
idf_t : Nilai Inverse Document Frequency

D. Naïve Bayes Classifier

Naïve Bayes Classifier merupakan metode yang sangat sering digunakan dalam data mining maupun teks mining, kemudahan dalam penggunaan metode ini adalah menjadi

alasan digunakannya metode ini[11]. Naïve bayes Classifier merupakan metode pengklasifikasian probabilistik berdasarkan teorema bayes dengan mempertimbangkan asumsi kemandirian naïve. Selain penggunaan metode yang mudah, akurasi yang didapat dari metode ini cukup akurat. Berikut persamaan umum dalam metode naïve bayes :

$$P(Cj|Wi) = \frac{P(Cj)\ x\ P(Wi|Cj)}{P(Wi)} \quad (2)$$

P(Cj|Wi) :Posterior, adalah kemunculan peluang pada kategori j tertentu ketika terdapat kemunculan kata i
P(Cj):Prior, adalah peluang kemunculan dokumen pada kategori j
P(Wi|Cj) : Likelihood atau Conditional Probability, adalah peluang sebuah kata i masuk ke dalam kategori j
P(Wi) : Evidence, adalah peluang kemunculan sebuah kata
i : indeks kata yang berawal dari 1 sampai dengan kata ke-k
j : indeks kategori yang berawal dari 1 sampai dengan kategori ke-n

Menghitung jumlah dokumen pada kategori tertentu digambarkan pada persamaan berikut:

$$P(C_j) = N(C_j) / N \quad (3)$$

N(Cj) : jumlah dokumen latih yang masuk dalam kategori j
N : jumlah keseluruhan dokumen

Multinomial Model merupakan model probabilitas yang penelti gunakan. Berikut merupakan persamaan Multinomial Model

$$\mathrm{P(w|c)} = \frac{Count\ (w,c) + 1}{Count(c) + |V|} \quad (4)$$

count(w,c) = jumlah kemunculan kata w pada kategori c
count(c) = jumlah total kemunculan semua kata pada kategori c
|V| = jumlah term unik atau fitur

Rencana Pengujian
Pengujian perhitungan akan dilakukan dengan 3 pengujian, yaitu :

a. Confusion Matrix
Confusion matrix atau error matrix adalah sebuah metode perhitungan akurasi terhadap sebuah sistem pada konsep data mining. Terdapat 4 istilah di dalam Confusion matrix yaitu, True Positif (TP), True Negatif (TN), False Positif (FP), dan False Negatif (FN).

b. Precision and Recall
Precision merupakan tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. Sedangkan Recall merupakan tingkat keberhasilan sistem menemukan kembali sebuah informasi.

c. Cross Validation
Cross Validation adalah metode statistik untuk mengukur kinerja model algoritma dimana data dipisahkan menjadi 2 subset, yaitu data latih dan data uji. Pada penelitian ini penulis menggunakan 10 k-fold cross validation.

HASIL DAN PEMBAHASAN

A. Input Data

Data yang dibutuhkan untuk melakukan sebuah percobaan yaitu dokumen yang mempunyai format csv. Dokumen yang di masukkan nantinya juga memiliki 2 kolom, yaitu kolom kalimat dan kolom kategori. Sebelumnya dokumen sudah diberikan kategori secara manual. Dokumen didapat dengan 2 cara, yaitu:

a. Crawling data dari twitter menggunakan RStudio dengan keyword Ibukota Pindah dan Ibukota Baru
b. Pencarian secara manual di twitter dengan keyword Ibukota Baru dan Ibukota Pindah dengan tujuan akun @jokowi, @TribunNews, @Kompascom, dan @pak_JK.

Dokumen yang didapat dari kedua jenis pencarian yaitu :
a. 123 Dokumen dengan Kategori Positif.
b. 173 Dokumen dengan Kategori Negatif.
c. 204 Dokumen dengan Kategori Netral.

B. Input Data

Preproses merupakan proses yang dilakukan sebelum proses perhitungan, fungsi dari proses ini adalah membersihkan data dari hal – hal yang tidak diperlukan, misalnya simbol, angka, spasi, kata yang tidak diperlukan, dan lainnya. Preproses ini secara umum dibagi menjadi 4 tahapan yang telah dijelaskan pada poin 2.2.

Pada penelitian ini peneliti menggunakan library sastrawi yaitu library yang berasal dari stemmer nadzief adriani yang telah dilakukan penyempurnaan. Library Sastrawi digunakan karena library ini merupakan library yang dibangun berdasarkan algoritma Nadzief dan Adriani yang mempunyai tingkat akurasi baik dalam stemming bahasa indonesia. Library ini juga telah mendapat perbaikan dari algoritma aslinya.

Pada penelitian ini peneliti juga memasukkan proses stopword removal yaitu proses menghilangkan atau menghapus kata yang terlalu banyak muncul atau tidak diperlukan dalam perhitungan, fungsi dari proses ini adalah untuk memperbesar presentase proses klasifikasi data uji. Berikut merupakan tabel daftar stopword yang telah penulis susun :

Tabel 1 Daftar Stopword

'gue', 'sedang', 'jadi', 'serta', 'tiap', 'pas', 'sih', 'kan', 'kita', 'siapa', 'biar', 'tahun', 'makin', 'aja', 'saja', 'demi', 'lah', 'bikin', 'apalagi', 'saja', 'memang', 'lain', 'supaya', 'para', 'karena', 'akan', 'kota', 'atau', 'bangun', 'sudah', 'harus', 'mana', 'tapi', 'nanti', 'sekarang', 'seperti',

'buat', 'rakyat', 'masih', 'lagi', 'sama', 'nya', 'bisa', 'bagaimana', 'belum', 'bukan', 'jangan', 'presiden', 'republik', 'perintah', 'kalimantan', 'lebih', 'baik', 'apa', 'jakarta', 'negara', 'indonesia', 'mau', 'joko', 'widodo', 'di', 'pindah', 'juga', 'itu', 'ada', 'dari', 'untuk', 'ini', 'pak', 'banyak', 'dengan', 'dan', 'kalau', 'ke', 'jadi', 'x', 'd', 'ya', 'yang', 'yg', 'tidak', 'tdk', 'gak', 'ibukota', 'baru', 'saya', 'aku', 'gua', 'gw', 'kamu', 'anda', 'lo', 'lu', 'loe', 'bas', 'kayak', 'ingin', 'benar',

C. Bagi Data

Data yang didapat dari twitter dibagi menjadi 2 subset yaitu data latih serta data uji. Pembagian dilakukan secara acak. Dokumen dibagi menjadi 60% data latih serta 40% data uji dari total 500 data. Pembagian data secara acak dilakukan guna mendapatkan nilai rata – rata akurasi yang didapatkan dari beberapa percobaan.

D. Proses Data Latih

Setelah pembagian data latih, proses selanjutnya adalah menghitung probabilitas pada data latih. Perhitungan ini dibagi menjadi beberapa sub berikut.

a. Perhitungan Term Frequency

Term frequency merupakan banyaknya kemunculan kata dalam satu dokumen. Karena fokus pada pembobotan pada penelitian ini adalah TF-IDF sehingga perhitungan *term frequency* hanya dilakukan dengan mencari kata per kata dalam sebuah kalimat tanpa harus menggunakan rumus.

b. Perhitungan TF-IDF

TF-IDF merupakan hasil perkalian dari nilai term frequency dan inverse document frequency. Nilai term frequency telah dibahas pada sub bab sebelumnya. Nilai inverse document frequency didapatkan dari log(seluruh dokumen/document frequency). Document frequency merupakan banyaknya dokumen dimana sebuah kata muncul. Sama halnya dengan term frequency, document frequency pun tak memerlukan rumus untuk menghitung nilainya, hanya dengan mengecek jumlah seluruh kata yang sama dan muncul pada berapa dokumen. Hasil yang penulis lampirkan pada tabel berikut hanya 5 baris awal dan 5 baris akhir dari seluruh jumlah baris dalam tabel TF-IDF.

Tabel 2 TF-IDF

idterm	term	tf	df	Idf = Log(d/df)	tf*idf
1	acara	1	2	Log(300/2) = 2,17319	1 * 2,17319 = 2,17319
2	taxiway	1	1	Log(300/1) = 2,47422	1 * 2,47422 = 2,47422
3	landas	1	1	Log(300/1) = 2,47422	1 * 2,47422 = 2,47422
4	udara	1	1	Log(300/1) = 2,47422	1 * 2,47422= 2,47422
5	tadi	1	1	Log(300/1) = 2,47422	1 * 2,47422 = 2,47422
2535	moga	1	12	Log(300/12) = 1,39504	1 * 1,39504 = 1,39504
2536	manfaat	1	4	Log(300/4) = 1,87216	1 * 1,87216 = 1,87216
2537	dapat	1	9	Log(300/9) = 1,51977	1 * 1,51977 = 1,51977
2538	rejeki	1	1	Log(300/1) = 2,47422	1 * 2,47422 = 2,47422
2539	barokah	1	1	Log(300/1) = 2,47422	1 * 2,47422 = 2,47422

c. Perhitungan NKelas

NKelas merupakan total nilai TF-IDF pada sebuah kata pada sebuah kategori. Misal kata “lahan” mempunyai nilai TF-IDF sebesar 1,69607 dan kata “lahan” muncul pada dokumen berkategori negatif sebanyak 6 kali. Sehingga nilai dari NKelas untuk kata “lahan” pada kategori Negatif yaitu 6 x 1,69607 = 10,17642. Sedangkan kata “lahan” tidak pernah muncul pada dokumen berkategori Positif dan Netral, sehingga nilai dari NKelas untuk kata “lahan” pada kategori Positif dan Netral adalah 0.

d. Perhitungan Probabilitas Kelas

Probabilitas kelas merupakan Multinominal Naïve Bayes, tahap ini adalah tahap akhir dari prose data latih, karena pada tahap ini telah didapat nilai probabilitas pada tiap kata di tiap kategori. Nilai ini yang nantinya dijadikan acuan sebagai perhitungan proses Data Uji.

Berikut Contoh perhitungan pada kata “acara” :

Positif :

Kata acara mempunyai nilai NKelas Positif sebesar 0.

Jumlah kata pada Kategori Positif yaitu 663.

Jumlah seluruh kata unik dalam data latih yaitu 1242.

Sehingga didapat perhitungan sebagai berikut :

$(0 + 1)/(663+1242) = 0{,}00052$

Sehingga nilai Probabilitas kata “acara” pada Kategori Positif yaitu sebesar 0,00052

Negatif :

Kata acara mempunyai nilai NKelas Negatif sebesar 0.

Jumlah kata pada Kategori Negatif yaitu 916.

Jumlah seluruh kata unik dalam data latih yaitu 1242.

Sehingga didapat perhitungan sebagai berikut :

$(0 + 1)/(916+1242) = 0{,}00046$

Sehingga nilai Probabilitas kata “acara” pada Kategori Negatif yaitu sebesar 0,00046

Netral :

Kata acara mempunyai nilai NKelas Netral sebesar 4,34638.

Jumlah kata pada Kategori Netral yaitu 960.

Jumlah seluruh kata unik dalam data latih yaitu 1242.

Sehingga didapat perhitungan sebagai berikut :

$(4{,}34638 + 1)/(960+1242) = 0{,}00243$

Sehingga nilai Probabilitas kata “acara” pada Kategori Netral yaitu sebesar 0,00243

Hasil yang penulis lampirkan pada tabel berikut hanya 5 baris awal dan 5 baris akhir dari seluruh jumlah baris dalam tabel Probabilitas Kelas.

Tabel 3 Tabel Probabilitas Kelas

term	Probabilitas Kelas		
	Positif	Negatif	Netral
acara	0.00052	0.00046	0.00243
taxiway	0.00052	0.00046	0.00158
landas	0.00052	0.00046	0.00158
udara	0.00052	0.00046	0.00158
tadi	0.00052	0.00046	0.00158
moga	0.01151	0.00046	0.00045
manfaat	0.00151	0.0022	0.0013
dapat	0.00451	0.00046	0.00322
rejeki	0.00182	0.00046	0.00045
barokah	0.00182	0.00046	0.00045

E. Data Data

Setelah dilakukan proses data latih, selanjutnya adalah proses pengujian data. Data uji merupakan 40% dari total 500 data yang telah dibagi secara acak pada tahap Bagi Data. Proses pengujian diawali dengan menghitung nilai prior. Nilai prior merupakan nilai kemunculan suatu dokumen dalam sebuah kategori., yaitu jumlah dokumen dalam sebuah kategori dibagi dengan seluruh dokumen data latih.

Prior Positif = 74/300 = 0,24666666666667

Prior Negatif = 97/300 = 0,32333333333333

Prior Netral = 129/300 = 0,43

Setelah prior didapat, barulah dokumen akan dipecah menjadi string, dan nilai probabilitas string akan dikalikan dengan nilai prior tersebut.

Hasil yang dicantumkan pada Tabel 4 merupakan hasil dari perhitungan 10 dokumen teratas pada data uji.

Tabel 4 Hasil Proses Data Uji 10 Dokumen

Kalimat	Positif	Negatif	Netral	Prediksi
gerindra bilangin prabowo menteri tahan tahan pulau lapis alutsista infrastruktur tahan langsung batas malaysia laut china selatan satu infantri tambah	4,034417385 3058E-44	2,419510559839 3E-45	1,706070436 4215E-43	Netral
acara taxiway landas udara tadi hendak inspirasi arsitek lupa masuk lapang raksasa	8,455683463 4406E-40	3,755694305678 6E-39	1,498307157 1131E-33	Netral
kosong besok najam pasir	5,792112213 3333E-14	5,667315598E-13	1,8055872E-13	Negatif
niat balik hindar praktik sembah berhala kronis	0,000128266 66666667	0,0010476	0,0001935	Negatif
rumah punya jamban rumah kolong jalan tol	1,33172798941 87E-20	8,4238670623333 E-20	5,404840 712064E-17	Netral
lari masalah gerindra ahok	8,60718666666 67E-7	5,38059E-7	2,554587 E-6	Netral
sayembaran gagas desain warganet sekali bulan agustus lalu umum ibu timur	8,66662806382 E-31	1,534773785933E -32	1,365800 6140257E -29	Netral
mega proyek hambalang mega proyek moga nasib hambalang	3,50821230733 65E-22	7,4945851702288 E-20	1,606778 6484375E -24	Negatif
agustus lalu cara resmi umum timur rencana sebut mulai	8,40818789019 02E-25	1,811566455061E -24	3,981530 4203167E -23	Netral
triliun hongkong demikian peras naik tagih bpjs	8,804224E-11	1,0292365517E-7	1,036192 5E-10	Negatif

F. Data Data

Dalam sebuah penelitian, tahap pengujian adalah hal yang sangat penting, tahap pengujian ini bisa menyimpulkan seberapa besar presentase dari metode yang digunakan dalam penelitian bisa berjalan. Dalam penelitian ini, penulis menggunakan 3 tahapan metode pengujian, yaitu :

a. Confussion Matrix

Confussion matrix atau *matrix error* ini digunakan untuk memasukkan menjabarkan nilai kebenaran dan nilai *error* yang dihasilkan dalam data uji ke dalam sebuah matrik. Pada umumnya *confussion matrix* terdiri dari *matrix* 2 x 2, namun pada penelitian ini, penulis menggunakan *matrix* 3 x 3 dikarenakan jumlah kategori yang dihasilkan dari proses klasifikasi pada penelitian ini berupa 3 klasifikasi. Berikut tabel *Confussion Matrix* :

Tabel 5 Confussion Matrix 3 x 3

		Kategori		
		Positif	Negatif	Netral
Prediksi	Positif	PP	NgP	NtP
	Negatif	PNg	NgNg	NtNg
	Netral	PNt	NgNt	NtNt

Baris Kolom yang berwarna biru merupakan klasifikasi yang terprediksi dengan benar dan kolom baris yang berwarna oranye terprediksi salah oleh sistem.

Keterangan :

PP = Kategori Positif yang terprediksi Positif (Benar).

NgP = Kategori Negatif yang terprediksi Positif (Salah).

NtP = Kategori Netral yang terprediksi Positif (Salah).

PNg = Kategori Positif yang terprediksi Negatif (Salah).

NgNg = Kategori Negatif yang terprediksi Negatif (Benar).

NtNg = Kategori Netral yang terprediksi Negatif (Salah).

PNt = Kategori Positif yang terprediksi Netral (Salah).

NgNt = Kategori Negatif yang terprediksi Netral (Salah).

NtNt = Kategori Netral yang terprediksi Netral (Benar).

Berikut merupakan tampilan dari Confussion matrix pada sistem

Prediksi	Kategori			
		Positif	Negatif	Netral
	Positif	23	8	7
	Negatif	15	59	17
	Netral	11	8	51

Gambar 1 Tampilan Confussion Matrix

Dari tampilan Gambar 1 dapat disimpulkan bahwa 23 dokumen terklasifikasi positif dengan benar, 59 dokumen terklasifikasi negative dengan benar, dan 51 dokumen terklasifikasi netral dengan benar. Selain itu dokumen terklasifikasi salah oleh sistem. Nilai diatas adalah acuan perhitungan dari metode pengujian selanjutnya, yaitu precision and recall.

b. *Precission and Recall*

Precision merupakan tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem.

Sedangkan Recall merupakan tingkat keberhasilan sistem menemukan kembali sebuah informasi.

Mengacu pada tabel Confussion matrix, perhitungan untuk precission serta recall pada matrixs 3 x 3 adalah sebagai berikut :

Precission Positif = PP/(PP+NgP+NtP) x 100% = 23/(23+8+7) x 100% = 60,5%

Precission Negatif = NgNg/(PNg+NgNg+NtNg) x 100% = 59/(15+59+17) x 100% = 64,8%

Precission Netral = NtNt/(PNt+NgNt+NtNt) x 100% = 51/(11+8+51) x 100% = 72,8%

Recall Positif = PP/(PP+PNg+PNt) x 100% = 23/(23+15+11) x 100% = 49,9%

Recall Negatif = NgNg/(NgP+NgNg+NgNt) x 100% = 59/(8+59+8) x 100% = 78,6%

Recall Netral = NtNt/(NtP+NtNg+NtNt) x 100% = 51/(7+17+51) x 100% = 68%

Akurasi Program = (PP+NgNg+NtNt)/(Jumlah Dokumen Data Uji) = (23+59+51)/200 = 66,83%

Berikut merupakan tampilan dari hasil precision and recall dan akurasi :

	Precision	Recall
Positif	60.526315789474	46.938775510204
Negatif	64.835164835165	78.666666666667
Netral	72.857142857143	68
Akurasi	66.834170854271	

Gambar 2 Tampilan Precission, Recall dan Akurasi

Hasil pada Gambar 2 akan menjadi acuan untuk menghitung metode pengujian yang terakhir, yaitu K-Fold Cross Validation.

c. *K-Fold Cross Validation*

Cross Validation adalah metode statistik untuk mengukur kinerja model algoritma dimana data dipisahkan menjadi 2 subset, yaitu data latih serta data uji. Pada penelitian ini penulis menggunakan 10 *K-Fold Cross Validation*, artinya penulis membagi data menjadi 2, yaitu data uji dan data latih yang dibagi secara acak dan untuk mendapatkan nilai pengujian dari metode uji ini adalah dengan melakukan 10 kali percobaan sehingga didapat nilai rata – rata akurasi dari sistem ini. Penulis telah melakukan 10 kali percobaan, dan berikut merupakan tabel akurasi sistem dari masing – masing percobaan.

Tabel 6 10 K-Fold Cross Validation

Percobaan Ke -	Nilai Akurasi
1	58.7 %
2	64.8 %
3	64.8 %
4	64.8 %
5	64.8 %
6	64.8 %
7	64.8 %
8	64.8 %
9	66.8 %
10	66.8 %

Untuk menghitung rata – rata akurasi maka menjumlahkan 10 hasil akurasi tersebut dan membaginya dengan 10.

Rata – rata = (58,7%+ 64,8%+ 64,8% + 64,8%+ 64,8%+ 64,8%+ 64,8%+ 64,8%+ 64,8%+66,8%+66,8%)/10

= 64,6%

Dari hasil ketiga metode uji didapatkan hasil akurasi program sebesar 64,6%.

KESIMPULAN

Dari penelitian yang sudah dilakukan dapat disimpulkan bahwa :

1. Prediksi dilakukan dengan menggunakan 3 kategori, yaitu Positif, Negatif dan Netral dengan menggunakan Metode Naïve Bayes Classifier dan Pembobotan TF-IDF
2. Penelitian dilakukan dengan membagi dokumen menjadi 2, yaitu Data Latih sebanyak 300 data dan Data Uji sebanyak 200 data dari Total 500 data.
3. Hasil perhitungan dari gabungan antara k-fold cross validation, Precision Recall, dan coffussion Matrix didapat hasil akurasi sebesar 64,6%.

REFERENSI

[1] W. L. Hutasoit, “Analisa Pemindahan Ibukota Negara,” Dedikasi, vol. 19, no. 2, pp. 108–128, 2018.

[2] A. Praditya, “Pengaruh Media Sosial Dan Komunikasi Bisnis Terhadap Perkembangan Bisnis Online Shop,” vol. 2, no. 1, 2019.

[3] W. Uriawan and H. Hidayat, “Rancang Bangun Aplikasi Pembelajaran Ilmu Sharaf Dalam Tata Bahasa Arab Berbasis Android,” ISTEK J. Kaji. Islam. Sains dan Teknol., vol. 10, no. 2, pp. 107–122, 2017.

[4] F. Nurhuda, S. Widya Sihwi, and A. Doewes, “Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier,” J. Teknol. Inf. ITSmart, vol. 2, no. 2, p. 35, 2016.

[5] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, “Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 12, pp. 1718–1724, 2017.

[6] D. G. Nugroho, Y. H. Chrisnanto, and A. Wahana, “Analisis Sentimen Pada Jasa Ojek Online ... (Nugroho dkk.),” pp. 156–161, 2015.

[7] N. Muchammad Shiddieqy Hadna, P. Insap Santosa, and W. Wahyu Winarno, “Studi [1] W. L. Hutasoit, “Analisa Pemindahan Ibukota Negara,” *Dedikasi*, vol. 19, no. 2, pp. 108–128, 2018.

[2] A. Praditya, “Pengaruh Media Sosial Dan Komunikasi Bisnis Terhadap Perkembangan Bisnis Online Shop,” vol. 2, no. 1, 2019.

[3] W. Uriawan and H. Hidayat, “Rancang Bangun Aplikasi Pembelajaran Ilmu Sharaf Dalam Tata Bahasa Arab Berbasis Android,” *ISTEK J. Kaji. Islam. Sains dan Teknol.*, vol. 10, no. 2, pp. 107–122, 2017.

[4] F. Nurhuda, S. Widya Sihwi, and A. Doewes, “Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier,” *J. Teknol. Inf. ITSmart*, vol. 2, no. 2, p. 35, 2016.

[5] A. Rossi, T. Lestari, R. Setya Perdana, and M. A. Fauzi, “Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa

Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1718–1724, 2017.
[6] D. G. Nugroho, Y. H. Chrisnanto, and A. Wahana, "Analisis Sentimen Pada Jasa Ojek Online ... (Nugroho dkk.)," pp. 156–161, 2015.
[7] N. Muchammad Shiddieqy Hadna, P. Insap Santosa, and W. Wahyu Winarno, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.
[8] C. Darujati, "Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa," *J. Link*, vol. 16, no. February 2012, p. 8, 2016.
[9] A. Fauzanu, E. Darwiyanto, G. Agung, A. Wisudiawan, F. Informatika, and U. Telkom, "Analisis Web Usage Mining Menggunakan Teknik K-Means Clustering Dan Association Rule (Studi Kasus : Www . Owlexa . Com) Web Usage Mining Analysis Using K-Means Clustering and Association Rule Technique (Case Study : Www . Owlexa . Com)," vol. 4, no. 2, pp. 3284–3291, 2017.
[10] M. A. Rosid, G. Gunawan, and E. Pramana, "Centroid Based Classifier With TF – IDF – ICF for Classfication of Student's Complaint at Appliation E-Complaint in Muhammadiyah University of Sidoarjo," vol. 1, no. 1, 2015.
[11] F. Rozy, S. Rangkuti, M. A. Fauzi, Y. A. Sari, E. Dewi, and L. Sari, "Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dengan Ensemble Feature dan Seleksi Fitur Pearson Correlation Coefficient," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 12, pp. 6354–6361, 2018.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relation- ships that could be construed as a potential conflict of interest.

Article History:
Received: 26-08-2020 | Accepted: 22-10-2020 | Published: 30-11-2020