REPOSITOR, Vol. 3, No. 5, November 2021, Pp. 493-500
ISSN : 2714-7975
E-ISSN : 2716-1382

Analisis Sentimen Pada Media Sosial Twitter Menggunakan *Naïve Bayes Classifier* Terhadap Kata Kunci "#Asiangames2018"

Abdy Yoga Syantara[*1], Evi Dwi Wahyuni[2], Vinna Rahmayanti Setyaning Nastiti[3]
[1,2,3]Universitas Muhammadiyah Malang
abdy.yoga@gmail.com[*1]

Abstrak

Data pengguna twitter di tahun 2016 ada 77% yang aktif dan 4,1 miliar tweet. Selain fungsi twitter sebagai media informasi, twitter juga memungkinkan setiap penggunanya menulis dan berbagi segala hal dengan menggunakan hashtag atau biasa dilambangkan dengan "#". Asian Games merupakan pesta olahraga Asia yang diselenggarakan setiap empat tahun sekali, dan melibatkan atlet-atlet dari seluruh Asia dimana acara bergengsi ini diorganizir oleh Olympic Council of Asia (OCA). Dengan begitu tujuan dari penelitian ini adalah untuk melihat sentiment masyarakat pada event Asian Games dengan memanfaatkan metode Naïve Bayes Classifier. Dataset yang diambil dari Twitter nantinya akan memasuki tahapan preprocessing. Hasil dataset tersebut nantinya akan dianalisa dengan metode Naïve Bayes Classifier dan menghasilkan prediksi nilai positif sebesar 27%, negatif sebesar 10% dan netral sebesar 16%. Dan hasil pengujian dengan akurasi sebesar 80,30%, serta precission, recall, dan f1-score masing masing sebesar 84%, 80%, dan 81%.

Kata Kunci: *Sentimen Analisis, Naïve Bayes, Asian Games 2018*

Abstract

Data on Twitter users in 2016 were 77% active and 4,1 billion tweets. In addition to the function of Twitter as a medium of information, Twitter also allows each user to write and share everything using a hashtag or usually denoted by "#". The Asian Games is an Asian sporting event which is held every four years, and involves athletes from all over Asia where this prestigious event is organized by the Olympic Council of Asia (OCA). Thus, the purpose of this research is to see the public sentiment at the Asian Games event by utilizing the Naïve Bayes Classifier method. The dataset taken from Twitter will later enter the preprocessing stage. The results of the dataset will be analyzed using the Naïve Bayes Classifier method and produce positive predictions of 27%, negative 10%, and neutral 16%. And the test results with an accuracy of 80.30%, as well as precision, recall, and f1-score respectively 84%, 80%, and 81%.

Keywords: *Analysis Sentiment, Naïve Bayes, Asian Games 2018*

1. Pendahuluan

Twitter merupakan salah satu sosial media hasil gagasan Dorsey yang kini melesat dan sangat diminati diberbagai belahan dunia [1]. Twitter menjadi salah satu sosial media yang sangat berkembang di abad 20. Dirilis oleh Twitter Indonesia data pengguna twitter di tahun 2016 ada 77% yang aktif dan 4,1 miliar tweet. Ini membuktikan kalau pengguna twitter di Indonesia terbilang cerewet dan masih berkisar umur remaja hingga dewasa. Dengan data diatas menguatkan argumen kita kalau Indonesia sebagai salah satu pengguna twitter terbanyak.

Selain fungsi Twitter sebagai media informasi, Twitter juga memungkinkan setiap penggunanya menulis dan berbagi segala hal. Baik itu informasi dibidang politik, pendidikan, hingga kesehatan. Setiap informasi yang ada di Twitter biasanya menggunakan *hashtag* atau biasa dilambangkan dengan "#". *Hashtag* ini berfungsi untuk memudahkan setiap orang mencari informasi, *hashtag* juga biasa digunakan dalam awalan kata kunci pada sebuah informasi. Setiap informasi yang memakai *hashtag* terbanyak biasanya akan menjadi trending topik atau tema pembahasan yang ada di Twitter pada suatu negara atau bahkan bisa menjadi trending topik pembahasan di dunia. Sebagai contoh tidak lama ini yang menjadi trending topik pada Twitter Indonesia adalah "#AsianGames2018". Pembahasan yang terdapat pada *hashtag* tersebut berisikan segala hal yang mengarah pada perhelatan pesta olahraga terbesar se-Asia yaitu Asian

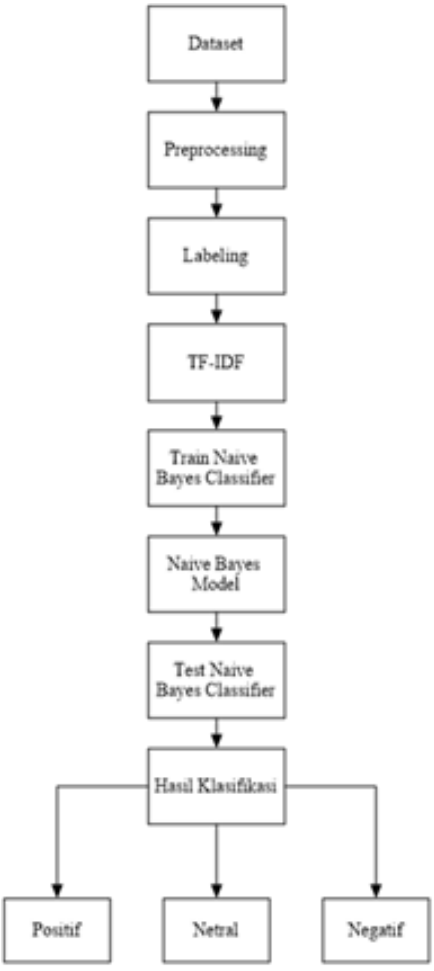
Games 2018 yang diadakan di Jakarta - Palembang. Dengan adanya pembahasan tersebut membuat #AsianGames2018 mencakup banyak informasi yang dapat disampaikan kepada masyarakat. Banyak dari masyarakat yang sangat antusias dengan adanya Asian Games, namun tidak sedikit pula yang mengkritik event tersebut. Untuk mengetahui sentimen yang terdapat dalam "#AsianGames2018", maka informasi tersebut haruslah diolah terlebih dahulu dengan metode klasifikasi.

Analisa sentimen merupakan studi komputasi mengenai pendapat, perilaku, dan emosi seseorang terhadap suatu entitas. Pada kasus ini sentimen analisis digunakan pada level data dokumen, orientasi ini dibagi menjadi tiga yaitu positif, negatif, dan netral. Perkembangan teknologi mendorong berbagai metode analisis untuk mempelajari dan mengklasifikasi dataset yang didapat dari sosial media. Salah satu metode yang dapat digunakan adalah Naïve Bayes Classifier atau NBC, Naïve Bayes Classifier adalah suatu metode klasifikasi yang sering digunakan karena mempunyai tingkat akurasi yang sangat tinggi. Pendekatan ini merupakan pendekatan yang mengacu pada Teorema Bayes yang merupakan prinsip peluang statistika untuk memecahkan masalah klasifikasi [2].

Adanya kondisi permasalahan pada media sosial Twitter yang telah dijabarkan sebelumnya. Maka, usulan penelitian ini adalah mendeteksi sentimen terhadap kata kunci #AsianGames2018 dengan menggunakan metode Naïve Bayes Classifier. Output penelitian ini adalah nilai sentimen untuk positif, negatif, dan netral. Serta menghitung akurasi, precision, recall, dan f1-score dari hasil klasifikasi.

2. Metode Penelitian

Ada beberapa tahapan yang dilakukan untuk analisis sentimen, semua tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1. Skema Implementasi

2.1 Dataset

Dataset yang akan digunakan pada penelitian ini berfokus pada data Twitter dengan kata kunci #AsianGames2018. Dengan data yang akan diambil dimulai dari tanggal 18 Agustus 2018 – 02 September 2018.

2.2 Preprocessing

Preprocessing merupakan suatu proses pembersihan data, reduksi data, dan diskritisasi data. Fase preprocessing tersebut akan membuat dataset lebih presisi. Dengan begitu digunakan tahapan *preprocessing* data untuk menyamaratakan data, menghapus kata yang tidak bermakna, mengeliminasi elemen dalam data, serta pengembalian kata awal [3]. Adapun proses – proses yang dilakukan pada tahapan *preprocessing* yaitu *case folding, tokenize, stopword, filtering,* dan *stemming.*

a. *Case Folding*

Case Folding adalah proses yang paling sering digunakan sebagai tahapan pertama dalam preprocessing. *Case Folding* bertujuan untuk menghapus semua huruf kapital atau huruf besar yang ada pada data dokumen, atau secara lebih gampang *Case Folding* mengembalikan semua huruf besar ke huruf kecilnya [4]. Sebagai contoh dari proses *Case Folding* ini dapat dilihat pada Tabel 1 dibawah.

Tabel 1. Contoh Case Folding

Teks	Hasil *Case Folding*
Semua kalangan dapat merasakannya, dimana Teknologi tersebut dapat mengubah pola pikir kalangan masyarakat saat ini.	semua kalangan dapat merasakannya, dimana teknologi tersebut dapat mengubah pola pikir masyarakat saat ini.

b. *Tokenize*

Tokenizing merupakan suatu proses normalisasi data tekstual yang sebelumnya kalimat yang kemudian di pecah menjadi kata perkata. Sebagai contoh dari proses *tokenizing* yang diterapkan pada suatu kalimat dapat dilihat pada Tabel 2 dibawah ini.

Tabel 2. Contoh Tokenize

Hasil *Case Folding*	Hasil *Tokenizing*
semua kalangan dapat merasakannya, dimana teknologi tersebut dapat mengubah pola pikir kalangan masyarakat saat ini.	‘semua’, ‘kalangan’, ‘dapat’, ‘merasakannya’, ‘dimana’, ‘teknologi’, ‘tersebut’, ‘dapat’, ‘mengubah’, ‘pola’, ‘pikir’, ‘kalangan’, ‘masyarakat’, ‘saat’, ‘ini’.

c. *Stopword*

Stopword merupakan suatu proses eliminasi kata yang bukan kata unik dari suatu dokumen. Seperti contoh kata ada, ini. Tabel 3 berikut contoh proses *stopword.*

Tabel 3. Contoh Stopword

Hasil *Tokenizing*	Hasil *Stopword*
‘semua’, ‘kalangan’, ‘dapat’, ‘merasakannya’, ‘dimana’, ‘teknologi’, ‘tersebut’, ‘dapat’, ‘mengubah’, ‘pola’, ‘pikir’, ‘masyarakat’, ‘saat’, ‘ini’.	‘semua’, ‘kalangan’, ‘merasakannya’, ‘teknologi’, ‘mengubah’, ‘pola’, ‘pikir’, ‘masyarakat’.

d. *Filtering*

Filtering merupakan tahapan eliminasi huruf yang bukan termasuk dalam a-z. Sebagai contoh proses *Filtering* sendiri dapat dilihat pada Tabel 4 dibawah ini.

Tabel 4. Contoh Filtering

Hasil *Stopword*	Hasil *Filtering*
‘semua’, ‘kalangan’, ‘merasakannya’, ‘teknologi’, ‘mengubah’, ‘pola’, ‘pikir’, ‘masyarakat’.	‘semua’, ‘kalangan’, ‘merasakannya’, ‘teknologi’, ‘mengubah’, ‘pola’, ‘pikir’, ‘masyarakat’.

e. *Stemming*

Stemming adalah proses penguraian berbagai bentuk atau variasi kata dari hasil tahapan sebelumnya untuk mengembalikan kata tersebut menjadi kata dasarnya (*stem*) seperti pada Tabel 5 [5].

Tabel 5. Contoh Stemming

Hasil *Filtering*	Hasil *Stemming*
‘semua’, ‘kalangan’, ‘ merasakannya’, ‘ teknologi’, ‘mengubah’, ‘pola’, ‘pikir’, ‘masyarakat’.	‘semua’, ‘kalangan’, ‘rasa’, ‘teknologi’, ‘ubah’, ‘pola’, ‘pikir’, ‘masyarakat’.

2.3 Labeling

Pemberian *labeling* digunakan untuk memudahkan proses – proses yang akan dilakukan, baik ditahap pelatihan maupun tahap pengujian. Label pada penelitian ini dibagi menjadi tiga kelas yaitu label positif, label netral, dan label negatif.

2.4 TF-IDF

TF-IDF merupakan sebuah algoritma yang dapat digunakan dalam proses pemberian bobot pada kata atau juga disebut perhitungan bobot *(W)*. *TF-IDF* yang digunakan dalam pembobotan kata berfungsi untuk mengekstraksi ciri dari suatu teks. Terdapat dua hal dalam perhitungan nilai bobot yaitu *TF* (*term frequency*) dan IDF (*inverse document frequency*). *TF* digunakan untuk mencari nilai dari kemunculan kata pada suatu dokumen, yang mana semakin banyak kata yang muncul maka semakin tinggi nilai dari *TF* tersebut. Sedangkan *IDF* adalah nilai kemunculan dari kata pada keseluruhan dokumen. Nilai *IDF* berbanding terbalik dengan nilai *TF*, semakin banyak kata yang muncul maka nilai dari *IDF* akan semakin kecil [6]. Nilai *IDF* sebuah kata (term) dapat dihitung menggunakan Persamaan 1 berikut.

$$\text{idf}_t = \log\left(\frac{N}{df_t}\right) \quad (1)$$

Dimana :

idf_t = Nilai *term IDF*.
N = Jumlah total semua dokumen atau jumlah total kalimat yang ada.
df_t = Jumlah dokumen yang mengandung *term* atau kata kunci.

Sedangkan algoritma yang digunakan untuk menghitung bobot (*W*) masing – masing dokumen terhadap sebuah kata kunci (*query*) menggunakan Persamaan 2 sebagai berikut.

$$W_{d,t} = tf_{d,t} * idf_t = \ tf_{dt}\ x\ log\left(\frac{N}{df_t}\right) \quad (2)$$

Dimana :

$W_{d,t}$ = Bobot kata *t* terhadap dokumen *d*.
$tf_{d,t}$ = Jumlah kemunculan *t* pada dokumen *d*.
N = Jumlah keseluruhan dokumen.
df_t = Jumlah dokumen yang mengandung *t*.

Dan nilai *tf* dapat diperoleh dari Persamaan 3 berikut.

$$tf_d = \frac{\text{Jumlah munculnya kata } t \text{ dalam dokumen}}{\text{Total jumlah seluruh kata dalam dokumen}} \quad (3)$$

Persamaan diatas menyesuaikan dengan tipe data dan tujuan dari penelitian.

2.5 Naïve Bayes

Naïve Bayes merupakan salah satu metode atau algoritma klasifikasi yang berakar pada teorema bayes, dimana proses klasifikasi dilakukan dengan cara melakukan perhitungan probabilitas dari suatu data [7]. Berikut Persamaan 4 *Naïve Bayes* yang digunakan dalam penelitian ini.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (4)$$

Dimana :

h = Hipotesis data yang merupakan suatu *class* spesifik.
D = Data dengan *class* yang belum diketahui
$P(h|D)$ = Probabilitas hipotesis *h* berdasarkan pada kondisi *D* (posteriori probabilitas)
$P(D|h)$ = Probabilitas *D* berdasarkan kondisi pada hipotesis *h*.
$P(h)$ = Probabilitas hipotesis *h* (prior probabilitas).
$P(D)$ = Probabilitas *D.*

Dimana *h* merupakan hipotesis data dari suatu *class*, sementara *D* merupakan data dengan *class* yang belum diketahui.

2.6 Pengujian

Pengujian pada penelitian ini dilakukan dengan menghitung *Accuracy, Precision, Recall,* dan *F1-Score* dari hasil klasifikasi sebelumnya yang ditampilkan pada *confusion matix.*

3. Hasil Penelitian dan Pembahasan

3.1 Dataset

Dataset yang digunakan berjumlah 665 tweet yang didapat dari hasil *crawling*, kemudian dari 665 dataset tersebut akan dibagi menjadi 80% data latih dan 20% data uji. Gambar 2 berikut adalah contoh data set yang akan digunakan.

Gambar 2. Dataset

3.2 Preprocessing Data

Proses ini singkatnya dilakukan untuk menyamaratakan data yang sudah didapat, membuang kata – kata yang tidak diperlukan, serta mengembalikan kata kebentuk dasarnya. Proses *preprocessing* yang digunakan dalam penelitian ini adalah *Case Folding, Tokenize, Stopword, Filtering, dan Stemming.*

3.3 Labeling

Pada tahap ini dataset akan diberikan label berdasarkan dengan kelasnya masing – masing, label yang diberikan dibagi menjadi tiga kelas, yaitu label positif, label netral, dan label negatif.

3.4 KFold Cross Validation

Kfold Cross Validation merupakan salah satu metode dari *Cross Validation* yang mempunyai fungsi untuk menilai suatu proses kinerja sebuah metode atau algoritma dengan membagi sampel data secara acak dan mengelompokan data tersebut. Kemudian salah satu data tersebut akan menjadi data uji, sedangkan sisa kelompok data yang lain akan menjadi data latih.

Nilai K yang digunakan pada penelitian ini adalah 5, dimana 5 sendiri merupakan nilai default yang diberikan dari *sklearn.* Nilai default value untuk K yang diberikan *sklearn* adalah 3 – 5 dan semakin kecil nilai K maka semakin kecil pula nilai akurasi yang didapatkan.

3.5 TF-IDF

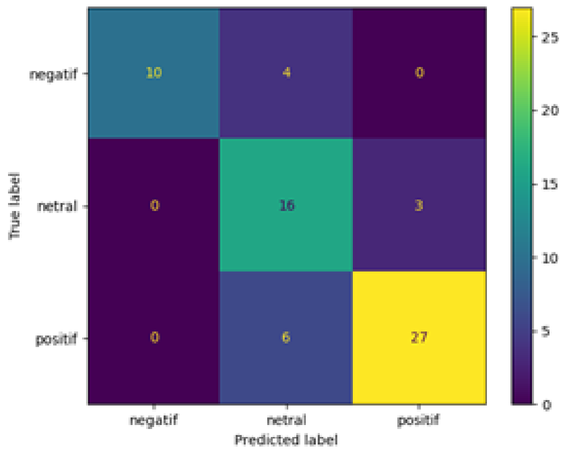
Tahapan kali ini akan dilakukan untuk mendapatkan nilai *TF* yang berupa frekuensi kemunculan kata dan *IDF* yang merupakan nilai dari setiap kata dalam keseluruhan dokumen. *TF-IDF* memberikat nilai bobot atau nilai *weight (W)* kepada setiap kata, dimulai dengan memecah kalimat ke kata. Hasil dari *TF-IDF* tersebut nantinya akan menjadi matrix angka seperti Tabel 6 berikut.

Tabel 6. Hasil TF-IDF

[[0.0, 0.16764547566435622, 2.2187979981117376, 1.741676743392075, 0.017400873791286103, 0.2743153259615689, 2.2187979981117376, 1.616738006783775, 2.2187979981117376, 2.0427067390560563, 0.42291798076766235, 2.2187979981117376, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.2743153259615689, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.7874342339527503, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.011972122079887832, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.18941422042652786, 0,

3.6 Klasifikasi Naïve Bayes

Dari hasil klasifikasi yang dilakukan dengan metode Naïve Bayes, data train akan digunakan terlebih dahulu dalam klasifikasi yang nantinya akan menjadi sebuah model. Setelah didapatkan model tersebut akan digunakan pada data uji dan memprediksi label dari data uji. Gambar 3 berikut hasil confusion matrix pada penelitian ini.



Gambar 3. Confusion Matrix

Dari hasil klasifikasi tersebut diketahui bahwa hasil prediksi dari positif positif adalah 27, sedangkan negatif negatif adalah 10, dan netral netral adalah 16. Serta akurasi yang diperoleh dari klasifikasi adalah 80,30%.

3.7 Akurasi

Dari Gambar 2 kita lihat nilai dari tiap label, dimana positif 27, negatif 10, dan netral 16. Dari situ kita dapat menghitung akurasi dari metode klasifikasi dengan menggunakan Persamaan 5 berikut.

$$Accuracy = \frac{(TPos + TNeg + TNet)}{(TPos + FPos + TNeg + FNeg + TNet + FNet)} \quad (5)$$

Dimana :
TPos = True Positif
TNeg = True Negatif
TNet = True Netral
FPos = False Positif
FNeg = False Negatif
FNet = False Netral

Dan diketahui bahwa nilai dari masing – masing seperti *TPos* = 27, *TNeg* = 10, *TNet* = 16, *FPos* = 6+0, *FNeg* = 4+0, dan *FNet* = 0+3. Dengan begitu didapatkan hasil akurasi sebesar 80.30%.

3.8 Precision, Recall, dan F1-Score

Nilai *precision, recall,* dan *f1-score,* yang didapatkan dari perhitungan sistem dapat dilihat pada Tabel 7 dibawah ini.

Tabel 7. Nilai setiap class

	Positif	*Netral*	*Negatif*
Precision	0,62	1,00	0,90
Recall	0,84	0,71	0,82
F1-Score	0,71	0,83	0,86

Dari hasil pengujian dengan pembagian data latih 80% dan data uji 20% dan n split yang dipilih adalah 5 maka diperoleh nilai seperti diatas. Dimana setiap nilai dari *class positif, class negatif,* dan *class netral* sangat stabil diatas angka 0,60. Nilai split juga berpengaruh pada hasil dari *precision, recall,* dan *f1-score,* dimana jika nilai split semakin kecil maka nilai akurasi, *precision, recall,* dan *f1-score* juga akan ikut menurun.

4. Kesimpulan

Dari hasil penelitian yang sudah dilakukan, maka dapat diambil kesimpulan saat pertama kali melakukan persiapan pengumpulan data, dilakukan dengan dua cara yaitu menggunakan bahasa pemrograman R dan menggunakan *Tweepy*. Kemudian diputuskan untuk menggunakan hasil *crawling* dengan *tweepy* dikarenakan lebih mudah pemahamannya serta data yang didapatkan lebih banyak. Kemudian agar data dapat diimplementasikan dengan maksimal, data tersebut harus terlebih dahulu melalui beberapa tahapan dalam *preprocessing* agar hasil yang didapat dari klasifikasi lebih akurat dan tepat. Analisis sentimen dengan menggunakan *Naïve Bayes Classifier* dimana pada penelitian ini menggunakan 665 data yang dibagi menjadi 80% data latih dan 20% data uji dengan hasil akurasi sebesar 80,30% dan nilai *positif* 27, *negatif* 10, dan *netral* 16. Dan terakhir dari hasil pengujian nilai K dengan range 3 – 5 didapatkan hasil paling optimal dan akurasi paling tinggi pada nilai K 5 dengan akurasi 80,30%, dibawah nya terdapat nilai K 4 dengan akurasi 67,07%, dan yang paling kecil dengan nilai K 3 dengan akurasi 49,09%. Dengan ini untuk kedepannya diharapkan pelenitian lain dapat menambahkan metode yang lebih serasi untuk disandingkan dengan klasifikasi sentiment, seperti contohnya metode *Social Network Analysis* yang bisa diterapkan pada studi kasus media sosial yang sama.

Referensi

[1] T. E. Damayanti, “Pemanfaatan Twitter sebagai Media Information Sharing di Perpustakaan (Studi Kasus Tentang Pemanfaatan Media Sosial Twitter Sebagai Media Information Sharing di Perpustakaan Wilayah Kota Surabaya),” *J. Airlangga Univ.*, vol. 3, no. 2, 2014.

[2] Y. T. Arifin, “Komparasi Fitur Seleksi Pada Algoritma Support Vector Machine Untuk Analisis Sentimen Review,” vol. 3, no. September, pp. 191–199, 2016.

[3] I. Rish, “An empirical study of the naive Bayes classifier,” *IJCAI-2001 Work. Empir. Methods AI (IBM Tech. Rep. RC22230)*, vol. 3, no. 22, pp. 41–46, 2001, [Online]. Available: https://sites.google.com/site/irinarish/publications/RC22230.pdf?attredirects=0&d=1&cm_mc_uid=80717800187114522513777&cm_mc_sid_50200000=1452251377.

[4] I. Zulfa and E. Winarko, “Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network,” *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 2, p. 187, 2017, doi: 10.22146/ijccs.24716.

[5] aisah rini Susanti, “Analisis klasifikasi sentimen twitter terhadap kinerja layanan provider telekomunikasi menggunakan varian naïve bayes aisah rini susanti,” 2016.

[6] A. Hidayat, “Impementasi Metode Term Frequency and Inverse Document Frequency Dan Marginal Relevance Untuk Monitoring Diskusi Online,” vol. 13, no. 2, pp. 151–159, 2016.

[7] V. S and J. R, “Text Mining: open Source Tokenization Tools – An Analysis,” *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.