Vol. 3, No. 2, Agustus 2022, hal. 152-158

p-ISSN: 2723-567X
e-ISSN: 2723-5661

Jurnal Computer Science and Information Technology (CoSciTech)

http://ejurnal.umri.ac.id/index.php/coscitech/index

K-nearest neighbor (KNN) untuk menganalisis sentimen terhadap kebijakan merdeka belajar kampus merdeka pada komentar twitter

Febby Apri Wenando*[1], Rahman Septiadi[2], Rahmad Gunawan [3], Harun Mukhtar[4], Syahril[5]
Email: [1]febby.apri@it.unand.ac.id, [2]150401055@student.umri.ac.id, [3] goengoen78@umri.ac.id, [4]harunmukhtar@umri.ac.id, [5]syahril@umri.ac.id

[1]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Aandalas
[2,3,4]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau
[5]Sistem Informasi, Fakultas Ilmu Komputer, Universitas Muhammadiyah Riau

Diterima: 7 Juli 2022 | Direvisi: 5 Agustus 2022 | Disetujui: 17 Agustus 2022

Abstrak

Kebijakan menteri Pendidikan dan Kebudayaan RI tentang Merdeka Belajar Kampus Merdeka dikeluarkan pada tanggal 11 Desember 2019 yang lalu menjadi perbincangan dan perdebatan dikalangan netizen twitter. Kebijakan ini juga sempat menjadi trending topic. Twitter merupakan media sosial yang memiliki banyak pengguna dan sangat populer dengan sarat komentar – komentar. Komentar pada twitter tidak hanya bernilai positif tetapi juga banyak yang bernilai negatif. Setiap kebijakan pastilah akan mendapat dukungan dan juga hambatan dari berbagai kalangan masyarakat. Penelitian ini dilakukan untuk mengtahui seberapa besar sentimen masyarakat yang bernilai positif maupun yang bernilai negatif. Penelitian ini menggunakan algoritma K-Nearest Neighbor (KNN). KNN digunakan untuk mengklasifikasi komentar positif dan komentar negatif sehingga diketahui seberapa besar tanggapan masyarakat pengguna twitter yang bernilai positif maupun negatif. Penelitian ini mengguanakan sample data sebanyak 700 data. Data tersebut dibagi menjadi 2 bagian yakni data latih dan data uji. Untuk data latih menggunakan 630 data data uji sebanyak 70 data. Setelah pengujian dilakukan didapatkan nilai pada k-8 sebesar 84,28 %. dan saat pengujian dilakukan menggunakan k-fold cross validation mendapatkan ilai sebesar 84,42 % pada fold=10. Hasil tersebut menunjukkan bahwa KNN sangat cocok digunakan untuk mengklasifikasi komentar pada twitter terutama kamus kebijakan penerintah tentang kampus merdeka.

Kata kunci: *merdeka belajar, klasifikasi, akurasi, K-Nearest Neighbor, k-fold cross validation*

K-nearest neighbor (KNN) to analyze sentiment toward the independent learning campus policy on twitter comments

Abstract

The policy of the Minister of Education and Culture of the Republic of Indonesia regarding Independent Learning of the Independent Campus was issued on December 11, 2019, which then became a conversation and debate among Twitter netizens. This policy has also become a trending topic. Twitter is a social media that has many users and is very popular with full of comments. Comments on Twitter are not only positive but also many are negative. Every policy will surely get support and also obstacles from various circles of society. This research was conducted to find out how much public sentiment is positive or negative. This study uses the K-Nearest Neighbor (KNN) algorithm. KNN is used to classify positive comments and negative comments so that it is known how many positive and negative twitter user responses are. This study uses a data sample of 700 data. The data is divided into 2 parts, namely training data and test data. For training data using 630 test data as much as 70 data. After the test was carried out, the value at k-8 was 84.28%. and when the test is carried out using k-fold cross validation, it gets a value of 84.42% at fold = 10. These results indicate that KNN is very suitable to be used to classify comments on Twitter, especially the government policy dictionary on independent campuses.

Keywords: *free to learn, classification, accuracy, k-nearest neighbor, k-fold cross validation.*

1. PENDAHULUAN

Perkembangan teknologi di Indonesia berkembang dengan pesat. Perkembangan teknologi informasi ini semakin memberi kemudahan kepada rakyat untuk mendapatkan informasi yang dibutuhkan. Kemajuan teknologi informasi ini membuat pengguna semakin asik dengan dunia barunya. Dunia baru tersebut dikenal dengan sebutan media sosial. Media sosial semakin ramai digunakan dan semakin merajai dunia. Masyarakat Indonesia baik yang di kota maupun yang didesa menggunakan media sosial. Pada tahun 2019 saja berdasarkan portal berita tribunnews.com mencatat bahwa Indonesia menduduki peringkat ke 4 setelah India, Amerika Serikat dan brazil[1]. Terdapat beberapa media sosial yang digemari masyaraat Indonesia diantaranya Twitter, facebook, dan lain-lain. Twitter merupakan media sosial yang paling banyak komentar sehingga pantas untuk diteliti. Nadiem Anwar Makarim selaku menteri pendidikan dan kebudayaan Republik Indonesia pada bulan desember 2019 mengeluarkan kebijakan merdeka belajar yakni terdiri dari 4 kebijakan diantaranya Ujian Sekolah Berstandar Nasional (USBN), Ujian Nasional (UN) dan Rencana Pelaksanaan Pembelajaran [2]. Perdebatan tentang kebijakan ini dibahas di twitter bahkan menjadi *trending topic* dikalangan netizen. Sentimen ungkapan *netizen* ini dapat di jadikan sebuah pengetahuan yang dapat di klasifikasi atau di kelompokkan dan dapat dihitung persentase terhadap isu merdeka belajar yang berkembang sebagai gambaran melihat respon masyarakat terhadap kebijkan ini. Dengan data yang berupa teks opini masyarakat di twitter, informasi ini dapat di olah menjadi sebuah pengetahuan terhadap topik merdeka belajar. Untuk menentukan apakah opini pengguna tersebut bersifat pro atau kontra, diperlukan sebuah sistem yang bisa menganalisis opini-opini tersebut secara otomatis atau disebut juga dengan sistem analisis sentimen. Algoritma kecerdasan buatan sangat cocok digunakan untuk melakukan klasifikasi tersebut[4].

Memahami, mengektrak dan mengolah data tektual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini disebut dengan sentimen analisis. Kecendrungan opini terhadap sebuah masalah atau oleh seseorang didapatkan dari menganalisis sentimen dari sebuah komentar secara tektual. Kecendrungan terdiri dari 2 opini yakni opini positif dan opini negatif. Hasil analisis opini dapat dijadikan sebagai pertimbangan untuk pengambilan keputusan. Jika opini masyarakat cenderung mengarah pada positif berarti kebijakan tersebut layak untuk dilanjutkan. Sebaliknya jika kecendrungan terhadap kebijakan tersebut justru mengarah pada negatif maka diperlukan suatu kebijakan untuk memberi pemahaman secara menyeluruh[3].Penelitian ini berfokus untuk mengklasifikasi *tweet* yang mengandung sentimen positif dan negatif terhadap dataset merdeka belajar menggunakan algoritma *K-Nearest Neighbor* (KNN). Dan menghitung akurasi yang dihasilkan oleh algoritma ini dengan membagi data dengan rasio 90 % digunakan sebagai data latih dan 10 % digunakan sebagai data uji, dan *K-Fold Cross Validation* digunakan untuk memvalidasi.

2. METODE PENELITIAN

Penelitian ini terdiri dari 5 tahapan pelaksanaan. Tahapan – tahapan ini terdiri dari pengumpulan data, pemrosesan data, pembobotan, pengujian, dan validasi. Gambar 1 menunjukan tahapan – tahapan dari proses selama penelitian berlangsung.

2.1. Tahap Pengumpulan Data

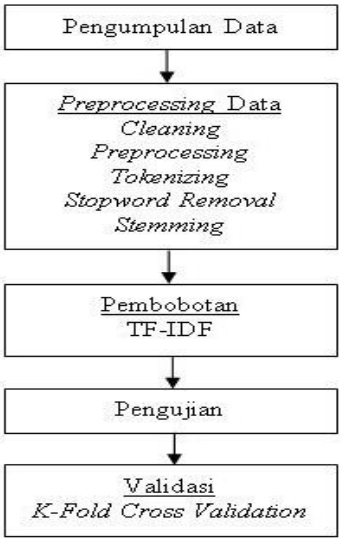
Data komentar dikumkumpukan menggunakan cara *crawling*. Hashtag #merdekabelajar, #menteripendidikan, #mendikbud, #nadiemmakarim diambil sebagai data contoh yang akan dibahas. Data yang sudah berhasil dikumpulkan diseleksi untuk mendapatkan yang yang baik sebanyak 700 data komentar sebagai data uji. Data tersebut selanjutnya disimpan dengan format csv. Data ini terdiri dari opini positif sebanyak 350 data dan data negatif terdiri dari 350 data.

2.2. Pemrosesan Data

Proses *crawling* digunakan untuk membersihkan data mentah agar dapat digunakan pada tahapan selanjutnya. Tahapan pemrosesan data Bertujuan untuk membersihkan data mentah yang didapatkan pada proses crawling agar siap digunakan pada tahapan selanjutnya, tahapan preprocessing data initerdiri dari 4 tahapan yakni pembersihan data, case folding, tokenizing, stopword removal dan steeming.

1) Pembersihan Data
   Tahapan pembersihan data dilakukan untuk mengurangi gangguan pada data. Pembersihan ini berguna untuk menghilangkan kata – kata yang tidak penting. Contoh kata yang tidak penting pada kasus ini seperti URL, *hashtag* (#), *username* (@username*), email*, *emoticon*, tanda baca seperti koma (,), titik (.) dan juga tanda baca lainnya.
2) *Case Folding*,
   Case folding digunakan untuk merubah huruf besar menjadi huruf kecil dan membuang huruf selain huruf abjat a sampai z. simbol ataupun angka semuanya akan dihilangkan sehingga sehingga hanya kata – kata penting saja yang digunakan[4]
3) Tokenizing
   Seluruh urutan karakter dipotong dan dijadikan satu potongan kata yang penting dan bisa digunakan dengan baik[5].
4) *Stopword Removal*
   Kata – kata yang sering muncul dan kata – kata yang tidak memiliki pengaruh akan dihilangkan. Hal ini digunakan untuk mencari ekstraksi sentimen yang benar. Kata petunjuk waktu dan juga kata tanya dihilangkan pada proses ini[6]
5) *Steeming*

Hasil *stopword removal* digunakan untuk mencari kata dasar (*stem*). Proses pencarian data dasar ini disebut dengan steeming. Aturan yang terdapat pada steeming terdiri dai 2 aturan yakni aturan dengan pendekatan kasus dan aturan dengan pendekatan aturan.



Gambar 1: Alur penelitian

2.3. Pembobotan Kata

Pembobotan sebuah kata yang dianggap sebanding dengan jumlah kemunculan kata pada sebuah dokumen disebut *Term Frequency* (TF). *Document Frequency* (DF) didapatkan dari TF. DF merupakan pemberian bobot untuk mengukur tingkat kepentingan sebuah kata dalam dokumen. Persamaan (1),(2), dan (3) digunakan untuk menghitung TF-IDF dengan mengkalikan TF dan IDF. IDF terbentuk dari hasil DF yang dibalik.

$$IDF(w) = log\left(\frac{N}{DF(w)}\right) \quad (1)$$

$$\text{Wdt} = \text{TF.IDF} \quad (2)$$

$$TF-IDF(w,d) = \frac{TF-IDF(w,d)}{\sqrt{\sum w=1n\, TF-IDF(w,d)^2}} \quad (3)$$

Dimana, *IDF(w)* = bobot kata dalam seluruh dokumen, *W* = sebuah kata, *TF(w,d)* = jumlah kemunculan kata *w* dalam kolom *d*, *IDF(w)* = *invers* DF dari kata *w*, *N* = jumlah seluruh dokumen, DF(w) = jumlah dokumen yang memuat kata *w*.

2.4. Klasifikasi *K-Nearest Neighbor*

Pengelompokan data baru berdasarkan jarak tetangga terdekatnya disebut KNN[7]. Kelas yang paling banyak muncul mengklasifikasikan hasil. Untuk mengklasifikasi teks dan data banyak peneliti yang menggunakan KNN. Metode Noural Network (NN) merupakan metode yang paling tua dan paling populer, metode NN digunakan sebagai dasar pengembangan terhadap KNN. Nilai K digunakan untuk menyatakan jumlah tetangga terdekat yang terlibat dalam penentuan prediksilabel kelas pada data uji. Voting kelas dari K dilakukan untuk memilih prediksilabe pada kelas data. Kelas dengan jumlah suara tetangga terbanyaklah yang diberikan sebagai label kelas hasil prediksi pada data uji tersebut [8].

Algoritma ini bertujuan untuk mengklasifikasikan objek menurut atribut dan sampel data latih [9]. Seperti halnya *K-means*, metode ini cukup sederhana, tidak ada asumsi mengenai distribusi data, mudah diaplikasikan dan sering dipakai dalam kasus nyata [7]. Di bawah ini adalah tahapan langkah algoritma K-NN:

1. Tentukan Nilai *k*
2. Hitung jarak data baru dengan semua data training
3. Urutkan jarak tersebut dari yang terdekat
4. Periksa kelas *k* tetannga tersebut
5. Kelas data baru = kelas mayoritas tetangga terdekatnya

Perhitungan K-NN adalah dengan menjumlahkan semua nilai kemiripan yang tergolong dalam satu kategori kemudian membandingkan manakah yang lebih besar. Persamaan K-NN adalah sebagai berikut :

$$cos\ Sim(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{(\sum_{i=1}^{n} A_i)^2}\sqrt{(\sum_{i=1}^{n} B_i)^2}} \quad (4)$$

Dimana : A = Data uji, B = Data latih, Ai dan B = bobot nilai yang diberikan untuk setiap *term* yang ada.

2.5. Pengujian Akurasi

Pengujian dilakukan untuk mendapatkan akurasi menggunakan *Confussion matrix*. C*onfussion matrix* digunakan untuk menghitung akurasi KNN yang digunakan untuk data hasil twitter tentang kampus merdeka. C*onfussion matrix* mengandung informasi yang membandingkan hasil pengklasifikasian secara otomatis[8]. Klasifikasi yang teridentifikasi oleh C*onfusion matrix* terdiri dari 2 output kelas yakni positif dan negative..

Tabel 1. *Confusion Matrix*

Kelas	Prediksi Positif	Prediksi Negatif
Aktual Positif	TP (*True Positive*)	FN (*False Negative*)
Aktual Negatif	FP (*False Positif*)	TN (*True Negatif*)

Dimana, TP = Data aktual positif, dan memprediksi positif, TN = Data aktual negative, dan memprediksi negatif, FP = Data aktual negative, dan memprediksi positif, FN = Data aktual positif, dan memprediksi negatif.

Confusion matrix digunakan untuk melakukan perhitungan dengan menghasilkan 4 output, diantaranya *accuracy*, digunakan untuk menghitung selisih antara data actual dengan data prediksi sebagaimana persamaan (5), *Precision* digunakan untuk menghitung tingkat ketepatan anatara informasi dengan jawaban yang didapat dari proses, sebagaimana persamaan (6). *Recall* digunakan untuk mengukur tingkat keberhasilan untuk menemukan Kembali sebuah informasi. Persamaan (7) digunakan untuk mencari *recall*. *Precision* dan *recall* jika digabungkan akan mendapatkan nilai F-measure, sebagai parameter tunggal menentukan ukuran keberhasilan. Persamaan (8) digunakan untuk menghitung nilai *F-measure* tersebut[10].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN} * 100\% \quad (5)$$

$$Precision = \frac{TP}{FP + TP} * 100\% \quad (6)$$

$$Recall = \frac{TP}{FN + TP} * 100\% \quad (7)$$

$$F_1 = \frac{Precision * Recall}{Precision + Recall} * 2 \quad (8)$$

2.6. K-Fold Cross Validation

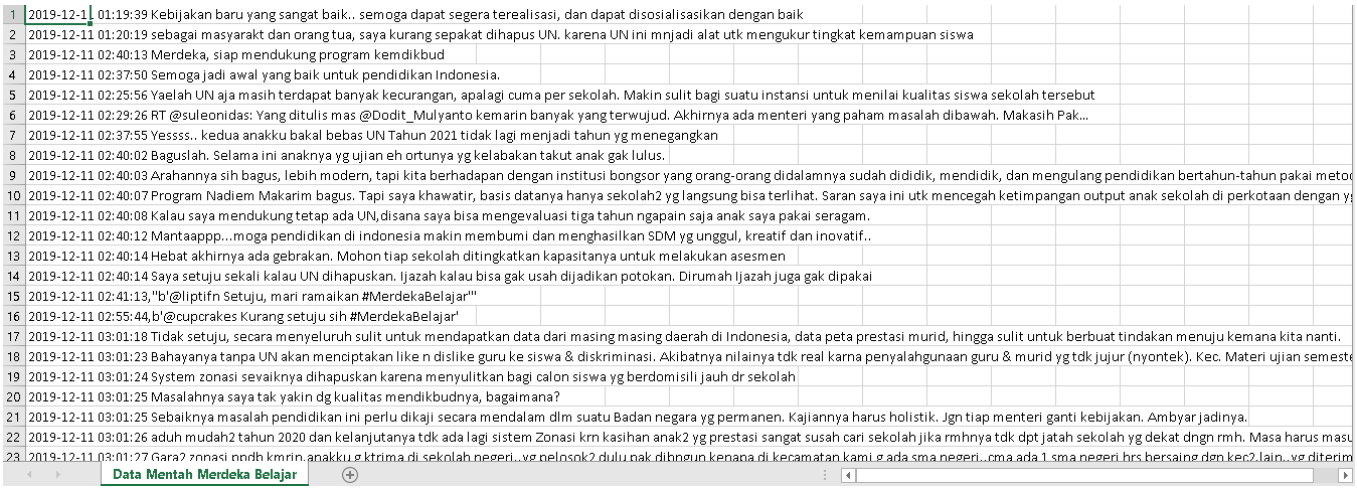
K-fold cross validation merupakan suatu cara yang digunakan untuk menguji kinerja metode klasifikasi. Cara yang digunakan untuk mencari kinerja tersebut dengan pengujian silang. Pengujian ini dialukan dengan membagi data menjadi dua bagian data yakni, data latih dan data uji. Pengujian dilakukan dengan proses berulang sampai pada tahap tertentu.[11]. C*ross validation* digunakan untuk memvalidasi metode untuk menilai hasil statistic dengan menggeneralisasi kumpulan data independent. Cara ini digunakan mengukur tingkat akurasi pada model prediksi saat diterapkan. cara yang digunakan pada *k-fold cross validation* dengan memecah data menjadi k bagian set data dengan ukuran yang sama. Fungsi *k-fold cross validation* untuk menghilangkan bias pada data. Pelatihan dan pengujian dilakukan sebanyak *k* kali[12].

3. HASIL DAN PEMBAHASAN

a. *Crawling* data

Penelitian ini memerlukan data, untuk mendapatkan data yang akan digunakan dengan melakukan pengumpulan data dengan menggunakan kata kunci pada twitter. Kata kunci yang dimasukkan pada proses *crawling* ini adalah #merdekabelajar, #menteripendidikan #mendikbud #nadiemmakarim sesuai dengan topik penelitian ini, data yang di *crawling* adalah *tweet* per tanggal 11 Desember 2019 – 18 Desember 2019, data yang sudah di*crawling* tersebut akan dipilih teks yang berbahasa Indonesia

sebanyak 700 dataset. data yang sudah didapatkan akan di konversi ke dalam format csv. Dataset dibagi kedalam 2 kelas yaitu positif sebanyak 350 data dan kelas negative sebanyak 350 data.



Gambar 2. Dataset

b. *Preprocessing*

Preprocessing bertujuan untuk membersihkan data mentah yang didapatkan pada proses *crawling* agar siap digunakan pada tahapan selanjutnya, berikut adalah proses *preprocessing*

Tabel 2. Proses Preprocessing

2019-12-11 01:19:39 Kebijakan baru yang sangat baik.. semoga dapat segera terealisasi,dan dapat disosialisasikan dengan baik.	
Tahap	**Hasil**
Cleaning	Kebijakan baru yang sangat baik semoga dapat segera terealisasi dan dapat disosialisasikan dengan baik
Case Folding	kebijakan baru yang sangat baik semoga dapat segera terealisasi dan dapat disosialisasikan dengan baik
Tokenizing	(kebijakan) (baru) (yang) (sangat) (baik) (semoga) (dapat) (segera) (terealisasi) (dan) (dapat) (disosialisasikan) (dengan) (baik)
Stopword Removal	(kebijakan) (baru) (sangat) (baik) (semoga) (dapat) (segera) (terealisasi) (dapat) (disosialisasikan) (baik)
Stemming	(bijak) (baru) (sangat) (baik) (semoga) (dapat) (segera) (realisasi) (dapat) (sosialisasi) (baik)

c. Klasifikasi dengan *K-Nears Neighbor*

Proses klasifikasi dilakukan menggunakan 700 dataset merdeka belajar yang telah di bobotkan, dimana 630 data dijadikan sebagai data training dan 70 data dijadikan sebagai data testing. Proses klasifikasi dilakukan pada *platform jupyter notebook* menggunakan Bahasa pemrograman *python*.

d. Pengujian

Pengujian dilakukan sebanyak 10 kali dengan mesukkan nilai k berbeda. Hasil pengujian nilai k dapat dilihat pada tebel berikut :

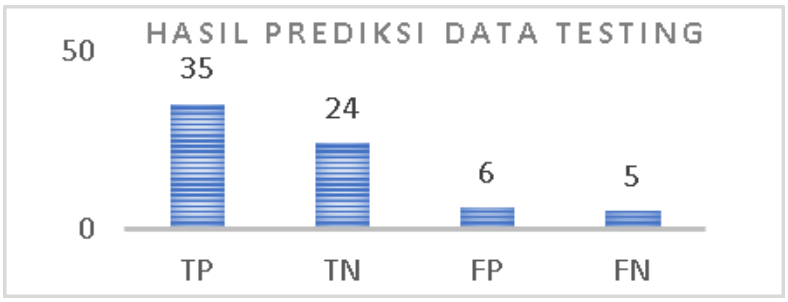
Tabel 3. Tabel Pengujian Nilai k

No	Nilai *k*	Accuracy	Precission	Recall	F-Measeure
1	1	80 %	80,4 %	80 %	80,2 %
2	3	78,57 %	78,74 %	78,57 %	78,65%
3	4	75,71 %	75,90 %	75,71 %	75,80 %
4	5	82,85 %	83,23 %	82,85 %	83,04 %
5	6	82,85 %	83,23 %	82,85 %	83,04 %
6	7	82,85 %	82,85 %	82,85 %	82,85 %
7	8	84,28 %	84,42 %	84,28 %	84,35 %

8	9	82,85 %	82,85 %	82,85 %	82,85 %
9	10	81,42 %	81,85 %	81,42 %	81,50 %
10	11	84,28 %	84,23 %	84,28 %	82,85 %

Hasil yang didapat setelah dilakukan proses klasifikasi menggunakan metode KNN menghasilkan nilai mendekati optimal. Nilai optimal ini dapat dilihat pada nilai k=8, dengan tingkat *accuracy* mencapai 84,28%, *precission* 84,42%, *recall* 84,28% dan *f-measure* 84,35%. Sedangkan nilai k yang menghasilkan nilai paling rendah yaitu k=15 dan tingkat *accuracy* terendah ada pada nilai k=4 yaitu 75,71%, *precission* 75,90%, *recall* 75,71%, *f-measure* 75,80%.

Setelah dilakukan pengujian maka didapatkan hasil data *testing* yang telah diklasifikasi oleh model *K-Nearest Neighbor* sebanyak 70 data dengan nilai TP 35 data, TN 24 data, FP 6 data dan FN sebanyak 5 data, hasil digambarkan dalam grafk berikut ini.



Gambar 3. Hasil prediksi

Metode KNN direkomendasikan untuk mengkalisifikasikan komentar twitter juga komentar lain yang sejenis. Selain difungsikan sebagai juga bisa direkomendasikan sebagai pemprediksi. Prediksi dapat digunakan membantu pengambilan keputusan oleh piahk menejemen[13].

e. Validasi

Folds Cross Validation digunakan pada penelitian ini untuk memvalidasi jumlah data latih dan data uji. Menentukan hasil *accuracy, precision, recall, dan f-measure* yang terbaik terbaik dari sebuah algoritma. Pengujian menggunakan *k-fold cross validation* dilakukan sebanyak 10 *fold*, yaitu 2,3,4,5,6,7,8,9,10,11 Hasil dari pengujian pengaruh nilai *fold* dapat dilihat pada tabel 5.

Tabel 5. Hasil Pengujian K-Fold Cross Validation

No	Nilai *fold*	Accuracy	Precission	Recall	F-Measeure
1	2	70,4 %	70,07 %	70,4 %	70,24 %
2	3	79,71 %	79,66 %	79,71 %	79,68%
3	4	82,14 %	82,08 %	82,14 %	82,11 %
4	5	83,57 %	83,50 %	83,57 %	83,53 %
5	6	83,14 %	83,08 %	83,14 %	83,11 %
6	7	83,28 %	83,22 %	83,28 %	83,25 %
7	8	83,14 %	83,09 %	83,14 %	83,11 %
8	9	84,14 %	84,09 %	84,14 %	84,85 %
9	10	84,28 %	84,23%	84,28 %	84,25 %
10	11	84,25 %	84,22 %	84,28%	84,25 %

Berdasarkan table 5 diatas dapat disimpulkan bahwa akurasi yang dihasilkan oleh k-fold cross validation mendapatkan nilai mendekali optimal yakni dengan nilai accuracy tertinggi adalah pada fold = 10 yakni 84,28 %, precission 84,23 %, recall 84,28 % dan f-measure 84,25. Akurasi terendah didapatkan pada nila fold = 2 yaitu 70,40 %, precission 70,07 %, recall 70,40 % dan f-measure 70,24.

4. KESIMPULAN

Berdasrakan hasil pengujian yang telah dilakukan pada penelitian ini maka dapat ditarik kesimpulan sebagai berikut :

1) Nilai akurasi tertinggi yang didapatkan dari proses sentimen analisis data komentar merdeka belajar menggunakan algoritma *K-Nearest Neighbor* dengan 700 dataset mengunakan perbandingan 90 % data *training*, 10 % data *testing* adalah pada nilai *k*-8 yaitu 84,28 %, *precission* 84,42 %, *recall* 84,28 % dan *f-measure* 84,35 %.
2) K-*fold cross validation* digunakan untuk memvalidasi dan menghasilkan nilai pada *fold*=10 mendapatkan *accuracy* tertinggi sebesar 84,28 %, *precission* 84,22 %, *recall* 84,28 % dan *f-measure* 84,25 %
3) Akurasi yang didapatkan pada pengujian menggunakan *k-fold cross validation* menghasilkan nilai yang hampir sama dengan pengujian yang menggunakan data 90 % data latih, dan 10 % data uji.
4) Berdasarkan hasil diatas metoda *KNN* cocok diterapkan untuk menganalsis sentimen dan mengklasifikasi dataset komentar merdeka belajar.

DAFTAR PUSTAKA

[1] C. Arifin, "Pengguna Sosial Media di Indonesia Terbesar Keempat di Dunia," *www.tribunnews.com*, 2019. https://www.tribunnews.com/techno/2019/06/19/pengguna-sosial-media-di-indonesia-terbesar-keempat-di-dunia?page=2 (accessed Apr. 01, 2020).

[2] M. Tohir, "Empat Pokok Kebijakan Merdeka Belajar," 2019, doi: 10.31219/osf.io/67rcq.

[3] F. J. D. Nitin Indurkhya, *Handbook of Natural Language Processing*, 2nd ed. CRC Press, 2010.

[4] J. S. Ronen Feldman, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007, 2007.

[5] R. Handoyo, R. Rumani, and S. M. Nasution, "Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K-Means Pada Pengelompokan Dokumen," *JSM STMIK Mikroskil*, vol. 15, no. 2, pp. 73–82, 2014.

[6] L. A. Utami, "Analisis Sentimen Opini Publik Berita Kebakaran Hutan Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization," vol. 13, no. 1, pp. 103–112, 2017.

[7] Santosa. B, *Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. 2011.

[8] E. Prasetyo, *Data Mining – Mengolah Data Menjadi Informasi Menggunakan Matlab*. ANDI, 2014.

[9] Moh Aziz Nugroho, "Dokumen Karya Ilmiah | Skripsi | Prodi Teknik Informatika - S1 | FIK | UDINUS | 2016," *Fik*, vol. 1, no. 1, pp. 1–2, 2016, doi: 10.1021/jf901375e.

[10] R. Hayami, Soni, and I. Gunawan, "Klasifikasi jamur menggunakan algoritma naïve bayes," *Jurnal Computer Science and Information Technology (CoSciTech)*, vol. 3, no. 1, pp. 28–33, 2020.

[11] R. Anand, K. Vishnu, and K. Burse, "K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 6, pp. 436–438, 2013.

[12] M. Brammer, *Principles of Data Mining*. Springer-Verlag London, 2007.

[13] M. Rifaldo, H. Mukhtar, R. M. Taufiq, and Y. Rizki, "Peramalan kedatangan wisatawan mancanegara ke indonesia menurut kebangsaan perbulannya menggunakan metode multilayer perceptron," *Jurnal Computer Science and Information Technology*, vol. 2, no. 1, pp. 113–119, 2021.