Komparasi Algoritma *Naïve Bayes* dan *k-Nearest Neighbor* Pada Klasifikasi Kontribusi Tokoh Politik

Moh. Ainur Rohman [1], Sri Harini [2,*]

[1] Magister Informatika; Universitas Islam Negeri Maulana Malik Ibrahim Malang;
Jl. Gajayana No.50, Dinoyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65144;
e-mail: 210605210005@student.uin-malang.ac.id.

[2] Magister Informatika; Universitas Islam Negeri Maulana Malik Ibrahim Malang;
Jl. Gajayana No.50, Dinoyo, Kec. Lowokwaru, Kota Malang, Jawa Timur 65144;
e-mail: sriharini@mat.uin-malang.ac.id.

* Korespondensi: e-mail: sriharini@mat.uin-malang.ac.id

Diterima: 09 Agustus 2022 ; Review: 08 September 2022; Disetujui: 05 Oktober 2022

Cara sitasi: Rohman MA, Harini S. 2022. Komparasi Algoritma *Naïve Bayes* dan *k-Nearest Neighbor* Pada Klasifikasi Kontribusi Tokoh Politik. Information System for Educators and Professionals. Vol 7(1): 21 – 30

Abstrak: Dalam berita politik, banyak sekali informasi tokoh-tokoh politik dalam mendongkrak elektabilitasnya. Berbagai kontribusi mereka lakukan seperti bidang pendidikan, infrastruktur, UMKM, kesehatan, teknologi, dan pelayanan publik. Oleh karena itu, perlu adanya klasifikasi berita menjadi beberapa kategori, tujuannya agar masyarakat mengetahui seberapa besar kontribusi para tokoh politik. Untuk mengatasi masalah tersebut, dibutuhkan sistem yang dapat mengkategorikan kontribusi-kontribusi para tokoh politik. Pada penelitian ini menggunakan dua algoritma untuk mengkomparasi algoritma mana yang terbaik untuk membangun sistem. Penelitian dilakukan menggunakan berbagai variasi jumlah dataset, dan tiga kali pengujian, untuk KNN dilakukan dengan 4 nilai k yaitu k=7, k=9, k=11 Hasilnya, algoritma KNN dengan k=7 yang terbaik dengan nilai *precision* sebesar 71.5%, nilai *recall* sebesar 22%, dan nilai *f-measure* sebesar 19.2%.

Kata kunci: *confusion matrix,* klasifikasi, *k-nearest neighbor, naïve bayes*

Abstract: In political news, there is a lot of information about political figures in boosting their electability. They made various contributions such as education, infrastructure, MSMEs, health, technology, and public services. Therefore, it is necessary to classify news into several, in order for the public to know how big the contribution category of political figures is. To overcome this problem, a system is needed that can categorize the contributions of political figures. In this study, two algorithms are used to compare which algorithm is the best to build the system. The study was conducted using various variations in the number of datasets, and three times of testing, for KNN carried out with 4 values of k, namely k=7, k=9, k=11. As a result, the KNN algorithm with k=7 is the best with a precision value of 71.5%, the recall value is 22%, and the f-measure value is 19.2%.

Keywords*: classification, confusion matrix, k-nearest neighbor, naïve bayes*

1. Pendahuluan

Berita adalah laporan yang memuat tentang kondisi, peristiwa dan situasi yang masih baru, penting dan menarik yang harus secepatnya dikabarkan atau disampaikan kepada publik [1].

Berita disajikan dalam berbagai bentuk seperti cetak, internet, siaran, dari orang ke orang dan masih banyak lagi. Namun seiring cepatnya perkembangan teknologi informasi, berita banyak disajikan melalui media digital yang ditayangkan dalam bentuk portal berita seperti kumparan, detik, vivanews, tribunnews, republika dan berita lainnya. Ada banyak kategori konten di portal berita online yang bisa dinikmati khalayak seperti hukum dan HAM, kesehatan, pariwisata dan budaya, sampai berita politik. Salah satu berita yang sering dikunjungi adalah berita politik. Hal ini dapat dibuktikan dari hasil survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2020 memberikan gambaran bahwa dari 16 kategori berita yang sering diakses, konten berita nomer tiga yang kerap kali diakses dengan persentase 7.8% [2].

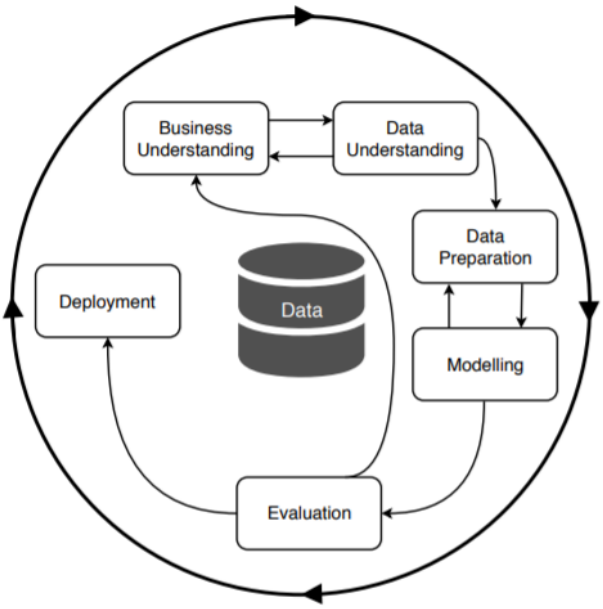
Di dalam konten berita politik banyak sekali informasi tokoh-tokoh politik mengenai bagaimana mereka mendongkrak elektabilitasnya, kegiatan kesehariannya, sampai seberapa besar kontribusi terhadap masyarakat. Berbagai kontribusi mereka lakukan seperti dalam bidang pendidikan, ekonomi, infrastruktur, teknologi, pelayanan publik, dan banyak lagi. Oleh karena itu, perlu adanya klasifikasi berita menjadi beberapa kategori, tujuannya agar masyarakat mengetahui seberapa besar kontribusi para tokoh politik.

Pada penelitian sebelumnya, Algoritma TF-IDF dan *cosine similarity* diterapkan pada penelitian klasifikasi teks berita online dengan skema empat pengujian [3], hasilnya dari rata-rata empat kali pengujian adalah 92.25 % dan hasil terbaik pada pengujian ketiga dengan data latih 70% dan data uji 30%. Klasifikasi juga digunakan untuk mengetahui penggunaan SMS terkait dan tidak terkait pekerjaan pada LKBN ANTARA [4], hasilnya presisi menunjukkan angka 96.02% sedangkan akurasinya 96.15%. Pada penelitian [5] menggunakan algoritma KNN untuk klasifikasi berita hoaks, hasil dari penelitian tersebut menunjukkan hasil yang tinggi, yaitu presisi 93.75% akurasi 92.31%, *recall* 90.90% dan *f1-score* 92.31%.

Berdasarkan penelitian sebelumnya, penulis mengusulkan untuk mengklasifikasikan kategori berita menggunakan algoritma Naïve Bayes dan KNN dengan metode penelitian CRISP-DM. Skema pengujian membandingkan algoritma Naïve Bayes dan KNN. Pengujian dilakukan tiga kali dengan komposisi data latih dan data uji 50:50, 70:30, dan 90:10. Untuk algoritma KNN menggunakan kombinasi k=7, k=9, dan k=11. Evaluasi pengujian menggunakan *confusion matrix* untuk menghitung akurasi, *recall,* dan *f1-score* nya. Sehingga akan diperoleh hasil terbaik dari perbandingan algoritma Naïve Bayes dan KNN untuk mengklasifikasikan kontribusi tokoh politik dari berita politik.

2. Metode Penelitian

Penelitian ini menggunakan metode *Cross-Industry Standard Process Model for Data Mining* (CRISP-DM). Menurut [6] dalam bukunya yang berjudul “CRISP-DM 1.0 Step-by-step data mining guide”, CRISP-DM adalah *blueprint* model proses dan metodologi *data mining* yang komprehensif. Ada enam tahap CRISP-DM, yaitu: 1). *Business Understanding,* 2). *Data Understanding,* 3). *Data Preparation*, 4). *Modeling*, 5). *Evaluation,* dan 6). *Deployment*.



Sumber: Martínez-Plumed, (2019)

Gambar 1. Tahapan CRISP-DM

1. *Business understanding*
   Pada tahap ini, dilakukan analisis tentang kontribusi apa saja yang sering dilakukan oleh para tokoh politik. Kontribusi para tokoh politik yang sering dilakukan berdasarkan berita online adalah di bidang Pendidikan, UMKM, Infrastruktur, Kesehatan, Teknologi, dan Pelayanan Publik.

2. *Data Understanding*
   Data understanding adalah tahap memeriksa data dan mencocokkan pada objek yang akan diteliti. Penelitian ini, dari empat atribut data yang diperoleh, hanya dua atribut yang akan digunakan. Yaitu Atribut berita dan kontribusi.

3. *Data Preparation*
   Untuk menghasilkan data yang berkualitas dan bisa diproses oleh algoritma *machine learning* perlu adanya persiapan data. *Data Preparation* menurut [7] ada tiga teknik, yaitu: *data cleaning, data integration,* dan *data reduction*. Penelitian ini menggunakan teknik *data cleaning* dengan cara: 1) *case folding,* 2) *tokenizing,* 3) *stopwords*, 4) *stemming.*

4. *Modelling*
   Pada tahap ini adalah membuat sebuah model prediksi. Model prediksi ini dibuat menggunakan algoritma *machine learning*. Pada penelitian ini menggunakan algoritma naïve bayes dengan teknik pembobotan TF-IDF:

   1. TF-IDF
      TF-IDF adalah suatu teknik memberikan bobot hubungan kata *(term)* terhadap dokumen. Teknik ini adalah sebuah ukuran statistik yang berfungsi untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen tunggal tiap kalimat. Berikut formula dari algoritma TF-IDF:

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

$$TF - IDF(w,d) = TF(w,d) \times IDF(w) \quad \text{..... (1)}$$

$IDF(w,d)$ = bobot suatu kata dalam keseluruhan dokumen
w = suatu kata (*word*)
d = suatu dokumen (*document*)
$TF(w,d)$ = frekuensi kemunculan sebuah kata *w* dalam dokumen *d*
$IDF(w)$ = *inverse* DF dari kata *w*
N = jumlah keseluruhan dokumen
$DF(w)$ = jumlah dokumen yang mengandung kata *w*

2. *Naïve Bayes*
Naïve bayes merupakan algoritma yang bekerja untuk menghitung setiap *term* yang ada dalam dokumen [8]. Dokumen dengan urutan kejadian yang muncul berdasarkan kata terhadap dokumen maka akan diabaikan, karena menjadi penyebab pengolahan kata menggunakan distribusi yang multinomial [9]. Berikut formula dari *Naïve Bayes:*

$$P(c|d) = P(c)\prod_{i=1}^{n} P(w_i|c) \quad \text{..... (2)}$$

d = besaran dokumen
n = jumlah semua kata yang ada pada dokumen

Selanjutnya nilai atau variabel P(c) diperoleh dengan formula berikut:

$$P(c) = \frac{N_c}{N} \quad \text{..... (3)}$$

P(c) = peluang kelas c
N = jumlah seluruh dokumen

Selanjutnya untuk menghitung peluang kata ke-*i* pada kelas *c* menggunakan formula berikut:

$$P(w_i|c) = \frac{count(w_i,c)+1}{count\ (c)+|V|} \quad \text{..... (4)}$$

$P(w_i|c)$ = Peluang kata ke-*i* pada kelas *c*
$count(w_i, c)$ = Jumlah kata ke-*i* pada kelas *c*
$count(c)$ = Jumlah semua kata pada kelas *c*
$|V|$ = Jumlah kata unik terhadap semua kelas

3. *K-Nearest Neighbor*
K-Nearest Neighbor (KNN) termasuk algoritma *supervised learning* yang digunakan untuk mengklasifikasikan objek berdasarkan *data training* yang memiliki jarak paling dekat dengan objek tersebut [10]. KNN bersifat sederhana, bekerja berdasarkan jarak terpendek dari *data testing* ke *data training* untuk menentukan kelas dari data tersebut setelah mengumpulkan data-data pada kelompok K tertentu, kemudian diambil kelas data mayoritas untuk dijadikan sebagai kelas prediksi dari *data testing* [4].

Ada banyak cara untuk mengukur jarak kedekatan antar data pada KNN diantaranya menggunakan *Euclidean distance. Euclidean distance* merupakan cara yang sering digunakan untuk menghitung jarak antar data. Jarak ini digunakan untuk menguji interpretasi kedekatan jarak antara dua objek. Berikut formula dari *Euclidean distance*:

$$d = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad \text{..... (5)}$$

d : Jarak
a : Data training/testing
b : Variabel data
n : Dimensi data

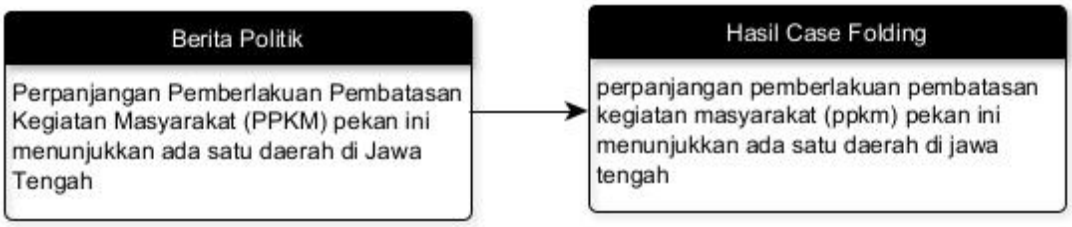
5. *Evaluation*
Pengujian model dilakukan untuk mengetahui kinerja pada algoritma yang diterapkan. Pengujian diukur dalam beberapa parameter. Parameter *accuracy, precision,* dan *f-measure*. Dan pengujian model menggunakan *confusion matrix. Confusion matrix* adalah suatu alat yang bertujuan untuk mengukur *accuracy* pada data mining [11].

3. Hasil dan Pembahasan

1. *Business Understanding*
Pada penelitian ini data didapat dari empat sumber berita online yaitu kompas, *tribunnews*, detik, dan JPNN. Data berasal dari lima kategori yaitu Pendidikan, Infrastruktur, UMKM, Pelayanan Publik, Teknologi, dan Kesehatan. Total datanya diperoleh sebanyak 525 artikel.

2. *Data Understanding* dan *Data Preparation*
Tahapan *data preparation* menggunakan 4 teknik *data cleaning* yaitu *Case Folding*, *Tokenizing, Stopwords,* dan *Stemming*.

   a. *Case Folding*
   Case folding merupakan tahapan mengkonversi semua karakter huruf pada dokumen menjadi huruf kecil. Pada proses ini karakter yang akan diproses hanya karakter alphabet yaitu huruf "a" sampai "z". Pada gambar 2 menampilkan artikel berita yang belum diproses sampai artikel diproses menjadi huruf kecil semua.



Sumber: Hasil Penelitian (2022)

Gambar 2. Proses *Case Folding*

b. *Tokenizing*
Tokenizing merupakan tahapan proses pemotongan kumpulan kata menjadi sebuah token. Pada tahapan ini spasi digunakan sebagai pemisah antar kata. Proses *tokenizing* dapat dilihat pada gambar 3.

Sumber: Hasil Penelitian (2022)

Gambar 3. Proses *Tokenizing*

c. *Stopwords*
Stopwords merupakan tahapan untuk menghilangkan kata yang tidak berhubungan atau tidak memiliki arti dengan subjek utama. Namun, meskipun kata yang tidak relevan dihapus tidak akan merubah maksud dari data. Gambar 4 menampilkan hasil dari proses *stopwords*.

Sumber: Hasil Penelitian (2022)

Gambar 4. Hasil *Stopwords*

d. *Stemming*
Stemming adalah tahap yang bertujuan untuk menghilangkan kata imbuhan seperti awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan kombinasi awalan dan akhiran (*confixes*). Proses *stemming* pada penelitian ini menggunakan *library* sastrawi. Pada gambar 5 menampilkan hasil dari proses *stemming*.

Sumber: Hasil Penelitian (2022)

Gambar 5. Hasil *Stemming*

3. *Modeling* dan *Evaluation*

1. Pengujian *Naïve Bayes*
Pengujian kinerja algoritma *naïve bayes* didasarkan pada perhitungan *confusion matrix* dengan menghitung nilai *precision, recall,* dan *f-measure*. Pengujian dilakukan sebanyak 3 kali dengan pembagian *dataset* 50:50, 70:30, dan 90:10. Hasil pengujian kinerja naïve bayes seperti pada tabel 1.

Tabel 1. Pengujian Naïve Bayes

Pengujian	Data latih	Data uji	*Precision*	*Recall*	*F-measure*
Pengujian 1	50%	50%	25%	25%	30%
Pengujian 2	70%	30%	22%	23.2%	30%
Pengujian 3	90%	10%	19.47%	20.5%	26.6%
Rata-rata			22.15%	22.9%	28.8%

Pada tabel 1 peneliti melakukan pengujian sebanyak tiga kali pengujian dengan komposisi data latih dan data uji adalah 50:50, 70:30, dan 90:10. Hasil pengujian tersebut menunjukkan bahwa semakin seimbang antara data latih dan data uji maka hasilnya semakin baik. Oleh karena itu, hasil pengujian terbaik pada pengujian 1 dengan *precision* 25%, *recall* 25%, dan *f-measure* 30% dengan rata-rata *precision* 22.15%, *recall* 22.9%, dan *f-measure* 28.8%

2. Pengujian *K-Nearest Neighbor*
   Pengujian algoritma selanjutnya adalah algoritma *K-Nearest Neighbor* (KNN) yang didasarkan pada perhitungan *confusion matrix* dengan menghitung nilai *precision, recall,* dan *f-measure*. Pengujian dilakukan sebanyak tiga kali dengan pembagian *dataset* seperti tabel 1 serta menggunakan empat nilai K yaitu k=7, k=9, k=11, dan k=13. Hasil pengujian kinerja KNN seperti pada tabel 2.

Tabel 2. Pengujian *k-nearest neighbor* dengan k=7

Pengujian	Data latih	Data uji	*Precision*	*Recall*	*F-measure*
Pengujian 1	50%	50%	71.5%	45%	49.2%
Pengujian 2	70%	30%	53.5%	41.7%	41.2%
Pengujian 3	90%	10%	43.3%	44.7%	42.6%
	Rata-rata		56.1%	37.13%	37.6%

Pada tabel 2 peneliti melakukan pengujian dengan variasi k=7 sebanyak tiga kali pengujian dengan komposisi data latih dan data uji adalah 50:50, 70:30, dan 90:10. Hasil pengujian tersebut menunjukkan bahwa semakin seimbang antara data latih dan data uji maka hasilnya semakin baik. Oleh karena itu, hasil pengujian terbaik pada pengujian 1 dengan *precision* 71.5%, *recall* 45%, dan *f-measure* 49.2% dengan rata-rata *precision* 56.1%, *recall* 37.13%, dan *f-measure* 37.6%.

Tabel 3. Pengujian *k-nearest neighbor* dengan k=9

Pengujian	Data latih	Data uji	*Precision*	*Recall*	*F-measure*
Pengujian 1	50%	50%	53.9%	42.2%	42.8%
Pengujian 2	70%	30%	53.9%	42.44%	42.8%
Pengujian 3	90%	10%	41%	42.7%	40.6%
	Rata-rata		49.6%	42.4%	42.06%

Pada tabel 3 peneliti melakukan pengujian dengan variasi k=9 sebanyak tiga kali pengujian dengan komposisi data latih dan data uji adalah 50:50, 70:30, dan 90:10. Hasil pengujian tersebut menunjukkan bahwa komposisi dengan data latih dan data uji 70:30 adalah yang terbaik. Oleh karena itu, hasil pengujian terbaik pada pengujian 1 dengan *precision* 71.5%, *recall* 45%, dan *f-measure* 49.2% dengan rata-rata *precision* 56.1%, *recall* 37.13%, dan *f-measure* 37.6%.

Tabel 4. Pengujian *k-nearest neighbor* dengan k=11

Pengujian	Data latih	Data uji	*Precision*	*Recall*	*F-measure*
Pengujian 1	50%	50%	36.7%	37.6%	36.6%
Pengujian 2	70%	30%	36.6%	37.04%	36.1%
Pengujian 3	90%	10%	38.8%	40.9%	39.1%
	Rata-rata		37.36%	38.21%	37.26%

Pada tabel 4 peneliti melakukan pengujian dengan variasi k=11 sebanyak tiga kali pengujian dengan komposisi data latih dan data uji adalah 50:50, 70:30, dan 90:10. Hasil pengujian tersebut menunjukkan bahwa semakin banyak data latih maka hasilnya semakin baik. Oleh karena itu, hasil pengujian terbaik pada pengujian 3 dengan *precision* 71.5%, *recall* 45%, dan *f-measure* 49.2% dengan rata-rata *precision* 56.1%, *recall* 37.13%, dan *f-measure* 37.6%.

Tabel 5. Hasil Rekapitulasi *Precision, Recall,* dan *F-Measure Naïve Bayes* dan KNN

		Naïve Bayes	KNN		
			K7	K9	K11
Pengujian 1	*Precision*	25%	71.5%	53.9%	36.7%
	Recall	22%	45%	42.2%	37.6%
	F-Measure	19.47%	19.2%	42.8%	36.6%
Pengujian 2	*Precision*	25%	53.5%	53.9%	52.9%
	Recall	23.3%	41.7%	42.44%	41.1%
	F-Measure	20.5%	41.2%	42.8%	40.7%
Pengujian 3	*Precision*	30%	43.3%	41%	38.2%
	Recall	30%	44.7%	42.7%	39.8%
	F-Measure	26.6%	42,6%	40,6%	38%

Pada tabel 5 menunjukkan nilai *precision, recall,* dan *f-measure* yang dihasilkan dari tiga kali pengujian dengan pengujian pertama menggunakan data *training* sebanyak 263 artikel dan data *testing* sebanyak 262 artikel. Pada pengujian ini ditunjukkan apabila menggunakan porsi data tersebut hasilnya untuk *naïve bayes* nilai *precision* sebesar 25%, nilai *recall* sebesar 22%, dan nilai *f-measure* sebesar 19.47%. Sedangkan nilai terbaik KNN ada pada k=7 dengan nilai *precision* sebesar 71.5%, nilai *recall* sebesar 22%, dan nilai *f-measure* sebesar 19.2%. Secara kumulatif nilai KNN yang terbaik dari *naïve bayes*.

Pada pengujian kedua menggunakan data *training* sebanyak 368 artikel dan data *testing* sebanyak 157 artikel. Pada pengujian ini ditunjukkan apabila menggunakan porsi data tersebut hasilnya untuk *naïve bayes* nilai *precision* sebesar 25%, nilai *recall* sebesar 23.3%, dan nilai *f-measure* sebesar 20.5%. Sedangkan nilai KNN terbaik ada pada k=9 dengan nilai *precision* sebesar 53.9%, nilai *recall* sebesar 42.44%, dan nilai *f-measure* sebesar 42.8%. Secara kumulatif nilai KNN yang terbaik dari *naïve bayes.*

Pada pengujian ketiga menggunakan data *training* sebanyak 472 artikel dan data *testing* sebanyak 53 artikel. Pada pengujian ini ditunjukkan apabila menggunakan porsi data tersebut hasilnya untuk *naïve bayes* nilai *precision* sebesar 30%, nilai *recall* sebesar 30%, dan nilai *f-measure* sebesar 26.6%. Sedangkan nilai KNN terbaik ada pada k=7 dengan nilai *precision* sebesar 43.3%, nilai *recall* sebesar 44.7%, dan nilai *f-measure* sebesar 42.6%. Secara kumulatif nilai KNN yang terbaik dari *naïve bayes*.

4. Kesimpulan

Dari hasil pengujian dan analisis nilai *precision, recall,* dan *f-measure* menggunakan kaidah *confusion matrix* untuk sistem analisis kontribusi tokoh politik menunjukkan nilai yang kecil dikarenakan *dataset* yang digunakan datanya *imbalanced*. Dan algoritma yang terbaik dalam klasifikasi kontribusi tokoh politik pada berita online adalah menggunakan kNN dengan nilai k=7.

Sistem klasifikasi terhadap kontribusi tokoh politik dibangun belum sempurna dan membutuhkan penyempurnaan pada penelitian lebih lanjut dikarenakan nilai *precision, recall,* dan *f-measure* yang masih kecil. Beberapa saran oleh penulis pada penelitian berikutnya sebagai berikut: Penulis menekankan optimasi pada data *imbalanced* dengan menggunakan metode *over-sampling* margin terlalu besarnya jarak antar metode. Dan metode *over-sampling* yang disarankan penulis menggunakan *repetition, bootstrapping,* dan SMOTE. Dan Penulis juga menyarankan pengujian menggunakan *K-fold Cross Validation*.

Referensi

[1] N. L. R. Maha Rani, “Persepsi Jurnalis dan Praktisi Humas terhadap Nilai Berita,” *J. ILMU Komun.*, vol. 10, no. 1, pp. 83–96, 2013, doi: 10.24002/jik.v10i1.155.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia, “Laporan Survei Internet APJII 2019 – 2020,” *Asos. Penyelenggara Jasa Internet Indones.*, vol. 2020, pp. 1–146, 2020, [Online]. Available: https://apjii.or.id/survei.

[3] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, “Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity,” *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 1, pp. 306–312, 2018, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/796.

[4] W. Gata, “Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS,” vol. 6, pp. 1–13, 2017.

[5] F. N. Rozi and D. H. Sulistyawati, “KLASIFIKASI BERITA HOAX PILPRES MENGGUNAKAN METODE MODIFIED K-NEAREST NEIGHBOR DAN PEMBOBOTAN

MENGGUNAKAN TF-IDF," *KONVERGENSI*, vol. 15, no. 1, Oct. 2019, doi: 10.30996/konv.v15i1.2828.

[6] P. C. Ncr *et al.*, "Crisp-Dm," *SPSS inc*, vol. 78, pp. 1–78, 2000, [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[7] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[8] S. Fanissa, M. A. Fauzi, and S. Adinugroho, "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.

[9] X. Feng, S. Li, C. Yuan, P. Zeng, and Y. Sun, "Prediction of Slope Stability using Naive Bayes Classifier," *KSCE J. Civ. Eng.*, vol. 22, no. 3, pp. 941–950, 2018, doi: 10.1007/s12205-018-1337-3.

[10] Henny Leidiyana, "Penerapan Algoritma KNN untuk Penentuan Resiko kredit Kepemilikan Kendaraan Bermotor," pp. 217–224, 2013.

[11] T. Rosandy, "Perbandingan Metode Naive Bayes Classifier Dengan Metode Decision Tree (C4. 5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus: Kspps/Bmt Al-fadhila," *J. Teknol. Inf. Magister*, vol. 2, no. 01, pp. 52–62, 2017.