Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 4, No. 1, Januari 2020, hlm. 373-380 http://j-ptiik.ub.ac.id

Klasifikasi Isu Suku, Antar Golongan, Ras, Agama (SARA) pada Twitter Berbahasa Indonesia menggunakan Metode *Improved K-Nearest Neighbor* (K-NN)

Firhad Rinaldi Saputra[1], Indriati[2], Sutrisno[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]firhadrinaldi22@gmail.com, [2]indriati.tif@ub.ac.id, [3]trisno@ub.ac.id

Abstrak

Twitter merupakan jejaring sosial yang memiliki pengguna aktif terbanyak saat ini. Pengguna Twitter berinteraksi dan berkomunikasi melalui pesan teks yang sering dikenal sebagai *tweet*. Pesan *tweet* yang dikirim oleh pengguna berisi tentang berita, sarana pembelajaran, kampanye politik, dan sentimen Isu Suku, Antar golongan, Suku, Agama (SARA). Pengguna twitter banyak yang menggunakan isu SARA sebagai topik pembicaraan sehingga berakibat salah penafsiran dari berbagai pihak. Pada sistem jejaring sosial twitter tidak bisa memeriksa *tweet* yang menggandung isu SARA, oleh karena itu dibutuhkan sistem cerdas untuk melakukan klasifikasi pada *tweet* yang memudahkan pengguna mengetahui isi *tweet* yang mengandung isu SARA atau bukan isu SARA. Proses klasifikasi menggunakan beberapa tahap proses yakni, *preprocessing* yang terdiri dari *cleansing, case folding, tokenisasi, filtering* dan *stemming* pada dokumen data latih dan data uji. Selanjutnya melakukan proses *term weghting*, hingga proses klasifikasi dengan metode *Improved K-Nearest Neighbor*. Pengujian pada penelitian Klasifikasi Isu Suku, Antar Golongan, Ras, Agama (SARA) pada Twitter Berbahasa Indonesia Menggunakan Metode *Improved K-Nearest Neighbor* (K-NN) dimana menggunakan 400 dokumen data mendapatkan hasil *Precision* sebanyak 0.976422, *Recall* sebanyak 1, *F-Measure* sebanyak 0.987944444, serta hasil Akurasi sebanyak 96%.

Kata kunci: *Text Mining, Klasifikasi, Isu Sara, Twitter, Improved K-Nearest Neighbor.*

Abstract

Twitter is a social network that has one of the most active users today. With the openness of information users move to send texts or tweets about other users, the number of Twitter users makes a lot of tweets related to ethnic issues, between groups, races, religions (SARA). Twitter cannot access the content of tweets that contain Sara's Issues, research is needed to classify tweets to understand including categories of Sara's Issues or Not Sara's Problems. Classification The Sara issue starts in several ways, namely preprocessing which consists of several stages, namely cleaning, folding cases, tokenisation, filtering and stemming. Followed by the term weghting process, to the classification process using the Improved K-Nearest Neighbor method. Based on the implementation and testing carried out in the research on Sara's Issue Classification on Twitter Using K-NN Increase, get the best results based on Precision averages of 0.976422, Remember at 1, F-Measure of 0.987944444 and Accuracy of 96%. Where the number of documents used as training data are 320 documents and test data as many as 80 documents. Where the number of documents, comparison or balance of training data and the value of k-value used determine the good or not classification process of the document.

Keywords: *Text Mining, Classification, Isu Sara, Twitter, Improved K-Nearest Neighbor*

1. PENDAHULUAN

Internet berkembang pesat setiap waktu yang menjadikan kebutuhan masyarakat dunia saat ini yang tidak bisa dipisahkan. Berkembangnya internet di Indonesia memudahkan pengguna untuk berkomunikasi satu sama lain, hingga menjadi kebutuhan gaya hidup saat ini. Pengguna internet Indonesia merupakan pengguna paling banyak situs jejaring media sosial sebesar 95 % pengguna aktif. (Bintang, 2013).

Situs jejaring media sosial sudah banyak dipakai oleh pengguna yang diataranya Twitter, Instragram, Facebook, Line, dan media sosial lainnya, media sosial membuat pengguna semakin memudahkan berintekaksi dengan pengguna lainnya. Pada tahun 2009 di mulailah media sosial twitter banyak disenangi oleh pengguna di Indonesia yang sampai saat ini mempunyai pengguna terbesar.

Pengguna yang menggunakan situs jejaring media sosial Twitter berinteraksi dengan pengguna lainnya dengan mengirimkan suatu pesan teks yang sering dikenal anak twitter yakni bernama tweet atau cuitan dalam bahasa gaulnya. Twitter dapat menghubungkan pengguna di seluruh dunia saling terhubung untuk memberikan pendapat maupun opini pribadi secara langsung (Xiong dan Liu, 2014). Sekarang banyak pesan teks yang berada di *timeline* twitter yang menyangkut pesan teks yang tidak jelas atau *spam* dan salah satu yang marak, yakni yang menyangkut isu suku, antar golongan, ras, agama (SARA).

Indonesia salah satu negara kepulauan terbesar didunia yang mempunyai kepulauan kecil hingga besar, dengan tersebarnya pulau-pulau beragaman pula penduduk yang mempunyai adat, suku, agama, golongan dan ras yang berbeda-beda setiap pulau. Perbedaan ini yang dahulunya menjadikan Indonesia kuat akan gotong-royongnya, sekarang berubah menjadi saling olok satu sama lain yang ditentangkan Isu Sara tidak mengacu pada perbedaan suku, ras, agama, namun disebabkan perbedaan mendasar tentang ekonomi, persaingan bisnis, pencapaian politik maupun pendidikan. Dari permasalahan tersebut, pengguna twitter menjadi kebingungan pada *timeline* yang mengandung Isu Sara maupun Bukan Isu Sara. Oleh sebab itu penelitian uji terhadap mengklasifikasi *tweet* Isu Sara Berbahasa Indonesia sangat dipermudah dengan menyaring *tweet* yang sesuai keinginan.

Klasifikasi ialah tahap-tahap proses mengumpulkan dan mengelompokkan sehingga menghasilkan nilai yang akurat pada setiap kelas-kelas kategori. Pengujian klasifikasi ini untuk membantu mengelompokkan teks tweet termasuk Isu Sara atau Bukan Isu Sara, maka algoritme Improved K-NN menjadi acuan yang baik dan benar (Megantara , 2010). Sehingga klasifikasi untuk metode algoritme Improved K-NN sangat akurat untuk mendapatkan kategori kelas yang tepat (Puspitasari et all, 2017).

Pada penelitian ini mengacu pada tiga penelitian yang menggunakan menggunakan metode algoritme *Improved KNN* untuk memproses klasifikasi pada dokumen, Pada penelitian (Nathania, 2017) rata-rata Precision bernilai 0,8946 dan Recall bernilai 0,9405, F-Measure bernilai 0,9155 dan hasil Akurasi bernilai 89,57%. Selanjutnya (Putri, 2013) rata-rata dari tahapan proses Precision bernilai 0,823, Recall bernilai 0,865, serta F-measure bernilai 0,843. Dan terakhir (Puspitasari, 2018) akhiran F-Measure tertinggi mendapatkan nilai sebesar 71,77%

Berdasarkan permasalahan maraknya Isu Sara diatas maka dalam penelitian ini menggunakan klasifikasi pada Isu Sara pada media sosial twtter yang menggunakan metode algoritme *Improved KNN* yang mendapatkan hasil akhiran jenis pelabelan Isu Sara maupun Bukan Isu Sara.

2. KAJIAN PUSTAKA

2.1 Isu Suku, Antar Golongan, Ras, Agama (SARA)

Isu Suku, Antar Golongan, Ras, Agama (SARA) adalah tindakan terhadap suatu identitas yang menyangkut agama, suku, ras maupun golongan yang di dalamnya terdapat perkataan yang tidak baik juga melibatkan kekerasan.

2.2 Twitter

Perkembangan jejaring sosial twitter yang saat ini yang semakin pesat sampai sekarang. Banyak pengguna Indonesia menggunakan media sosial twitter untuk mencari informasi-informasi terkini, sepeti musik, seputar olahraga, seputar politik, hiburan, kemaceran jalan, seputar tempat liburan hingga berita terkini (Twitter, 2019).

2.3 *Preprocessing*

Preprocessing merupakan tahapan dokumen data yang diubah menjadi data yang tersusun kelas kategori yang diinginkan. Tahapan proses perhitungan *preprocessing*, yakni *cleansing*, *case folding*, *tokenizing*, *fitering*, dan *stemming.*

1. C*leansing*

Pengujian tahap awal pemrosesan dokumen data yang dihapus karakter huruf yang tidak digunakan seperti hashtag (#) dan username (@) serta menghapus link dan angka (Nur dan Santika., 2011).

2. *Tokenizing*

Tokenizing merupakan tahapan dokumen data yang dipisahkan oleh whitespace yang menghasilkan token-token pada dokumen (Nathania dkk, 2017).

3. *Filtering*

Filtering merupakan tahapan dokumen data diambil kata inti dan menghapus kata yang tidak diperlukan (Nurul, 2018).

4. *Stemming*

Stemming ialah tahap menguji dokumen data yang diubah menjadi kata dasar keseluruhan.

2.4 Pembobotan Kata

1. Term Frequency (TF)

yakni sekumpulan token bermunculan ada dalam dokumen data dan dibedakan banyaknya setiap dokumennya, yang diwujudkan persamaan 1.

$$W_{tf\ (t,d)} = 1 + \log TF\ (t,d) \quad (1)$$

Keterangan:

$W_{tf_{t,d}}$ = Banyak muncul kata di dokumen

2. Inverse Document Frequency (IDF)

Idf ialah persamaan yang memproses sejumlah dokumen data yang memiliki kata, diwujudkan hitung persamaan 2.

$$\text{idf}_t = \log_{10} N/\text{df}_t \quad (2)$$

Keterangan:

df_t = jumlah data

N = Jumlah keseluruhan data

3. TF-IDF Weighting

Tf-Idf Weighting ialah tahap mengolah proses pembobotan pada term yang dihasilkan oleh perhitungan yang mengalikan proses tf dan proses idf, yang diwujudkan persamaan 3 dan persamaan 4.

$$W_{t,d} = W_{tf_{t,d}} \text{ x idf}_t \quad (3)$$

Normalisasi:

$$W_{t,d} = \frac{W_{t,d}}{\sqrt{\sum_{t=1}^{n} W_{t,d}{}^2}} \quad (4)$$

4. Cosine Similarity

Menurut Adikara. (2017) tahap cosine similarity untuk membandingkan term yang ada dengan dokumen untuk mendapatkan hasil kemiripan yang di inginkan, yang diwujudkan persamaan 5 dan 6.

Tanpa normalisasi Wt,:

$$\text{CosSim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \cdot |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t}(W_{ij} \cdot W_{iq})}{\sqrt{\sum_{i=1}^{t} W_{ij}{}^2 \cdot \sum_{i=1}^{t} W_{iq}{}^2}} \quad (5)$$

Melalui proses menormalisasi yang berdasar pada persamaan sebelumnya:

$$\text{CosSim}(d_j, q) = \vec{d}_j, \vec{q} =$$

$$= \sum_{i=1}^{t}(W_{ij} \cdot W_{iq}) \quad (6)$$

Keterangan:

dj = Nilai data latih

q = Nilai data latih tetangga

Wij = Persamaan dokumen latih

Wiq = Persamaan tetangga dokumen latih

2.5 Improved K-Nearest Neighbor

Pengolahan metode algoritme *Improved K-NN* yang membedakan antara metode K-NN biasa adalah dalam menentukan awal nilai n k-values, dalam Improved K-NN penentuan nilai k-values dapat dimodifikasi sesuai keinginan, namun sebaliknya dalam metode K-NN untuk menentukan nilai awal k-values harus dalam keadaan terbaik. Keunggulan dari metode Improved K-NN tidak terpengaruh data latih yang besar.

Menentapkan nilai *k* baru(n).

$$n = \frac{k * N(c_m)}{Maks[N(c_m)|j=1..N_c]} \quad (7)$$

Keterangan:

- n = nilai *k* awal yang terbaru
- k = nilai *k* yang tetap
- $N(c_m)$ = nilai kemunculan frekuensi yang banyak pada dokumen
- $Maks[N(c_m)|j = 1..N_c]$ = semua kelas kategori yang memiliki data latih yang terbanyak

Setelah itu menentukan hasil proses kelas kategori yang menghitung banyaknya data latih dikalikan dengan data uji sebanyak jumlah nilai n tetangga.

$$p(x, c_m) = argMaks_m = \quad (8)$$

$$= \frac{\sum d_j \in top\_n\_kNN\ (c_m) sim(x, d_j) y(d_j, c_m)}{\sum d_j \in top\_n\_kNN\ (c_m) sim(x, d_j)}$$

Keterangan:

- $p(x, c_m)$: nilai dokumen X anggota c_m
- $sim(x, d_j)$: nilai sama antar dokumen X dokumen latih d_j

- top_n_kNN : munculnya nilai n terbaik di tetangga
- $y(d_j, c_m)$: nilai kelas kategori yang sudah memenuhi kelas kategori

2.6 Evaluasi

Proses tahapan evaluasi yakni menguji dan memeriksa tingkat akurasi dan keakuratan pada tingkat kelas kategori.

Tabel 1 *Confusion Matrix*

Hasil Prediksi	Hasil Aktual	
	SARA	Fakta
Isu Sara	TP	FP
Bukan Isu Sara	FN	TN

Keterangan:

- *True Positive* (TP), yakni sejumlah data yang benar sesuai kelas setiap kategori diprediksi benar oleh sistem.
- *True Negative* (TN), yakni sejumlah banyaknya data yang salah sesuai kelas setiap kategori diprediksi benar oleh sistem.
- *False Positive* (FP), yakni sejumlah banyaknya data yang salah kelas kategori diprediksi salah oleh sistem.
- *False Negative* (FN), yakni sejumlah banyaknya data yang benar kelas kategori diprediksi salah oleh sistem.

1. *Precision*

Precision merupakan memproses klasifikasi yang mendapatkan hasil ketepatan sesuai kelas kategori aslinya.

$$Precision = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \quad (9)$$

2. *Recall*

Recall merupakan memproses klasifikasi yang mendapatkan hasil kesesuaian kelas kategori dan mengenalinya,

$$Recall = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (10)$$

3. *F-Measure*

Measure ialah proses perhitungan mengalikan *recall* dan *precision* yang selanjutnya dikalikan 2, yang menghasilkan kesesuaian dengan sistem.

$$\mathrm{F} = (2 \times \mathrm{P} \times \mathrm{R})/(\mathrm{P}+\mathrm{R}) \quad (11)$$

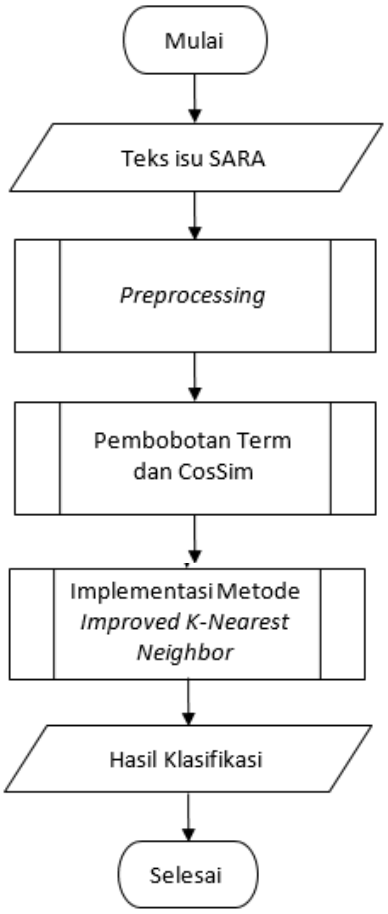
4. Akurasi

Akurasi merupakan tahapan membandingkan nilai proses prediksi yang sudah dihitung pada pengujian nilai aktual.

$$\text{Akurasi} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} * 100 \quad (12)$$

3. METODOLOGI PENELITIAN

Penelitian Isu Sara ini memakai data primer, yaitu kumpulan data teks *tweet* berbahasa Indonesia yang dikumpulkan sebanyak 400 dokumen data. Pada teknik pengumpulan dan pengelompokan dokumen data yang berada di twitter yang terindikasi kedalam kategori Isu Sara.



Gambar 1 Diagram Alir Sistem Klasifikasi Isu SARA pada Twitter Berbahasa Indonesia Mengunakan Metode *Improved KNN*

1. Teks Isu Sara

Memasukkan dokumen data tweet ke dalam sistem sehingga mendapatkan hasil akhiran kelas kategori Isu Sara atau Bukan Isu Sara.

2. *Preprocessing*

Preprocessing yakni mengolah dokumen

data tweet sehinngga mendapatkan data yang terstruktur. Pada pengolahan Text preprocessing dilakukan dengan beberapa tahap, diantaranya *cleansing, case folding, tokenisasi, filtering* dan *stemming.*

3. Perhitungan TF-IDF

Pada proses menghitung awal pembobotan menggunkan TF-IDF yang nantinya menghasilkan nilai unik, dan selanjutnya dilakukan persamaan nilai Wt,d yang sudah dinormalisasi.

4. Pembobotan Term dan Cosine Similarity

Setelah mendapatkan nilai proses persamaan perhitungan (Wt,d), Nilai cosine similarity dan nilai Wt.d dipadukan dan dihitung yang selanjutnya menghasilkan kelas kategori.

5. Klasifikasi Improved K-Nearest Neighbor

Selanjutnya melakukan menghitung n(k-vales baru) pada kelas kategori. Setelah mendapatkan kelas kategori maka menghitung probabilitas data uji pada kelas kategori.

6. Hasil Klasifikasi

Dan setelah semua tahap telah diselesaikan, sehingga menghasilkan kelas kategori Isu Sara atau Bukan Isu Sara.

4. HASIL DAN PEMBAHASAN

Proses yang diujikan pada setiap scenario memakai data dokumen sebanyak 400 dokumen. Banyak jumlah data latih dan data uji berada dalam 400 dokumen tersebut, yang di dalamnya terisi data dokumen latih sebanyak 320 data maupun terisi data dokumen uji sebanyak 80 data. Selanjutnya dokumen data latih maupun dokumen data uji dilakukan perhitungan pengujian skenario bergantian hingga 400 data dokumen terpenuhi, sehingga menghasilkan n(k-values baru), precision, recall, f-measure dan akurasi.

4.1 Pengujian Skenario

Diawal proses pengujian skenario dokumen data latih dan n k-values baru memiliki jumlah data yang tidak satu sama lain. Pada dokumen data latih menggunakan 320 data maupun dokumen data uji menggunakan 80 data yang digunakan sebagai beberapa faktor yang dipengaruhi data uji dan k-values awal terhadap proses pengklasifikasi. Pada tabel 2 dibawah ini merupakan perhitungan pengujian pada setiap masing-masing skenario.

1. Hasil Pengujian Skenario 1

Tabel 2 Hasil Pengujian Skenario 1

k-values	n (k values baru)		Precision	Recall	F-Measure	Akurasi
	Isu SARA	Bukan Isu SARA				
2	2	1	1	1	1	100%
4	4	2	1	1	1	100%
6	6	4	1	1	1	100%
8	8	5	1	1	1	100%
10	10	6	1	1	1	100%
15	15	9	0.9803	1	0.99	98%
20	20	12	0.9803	1	0.99	98%
25	25	15	0.9615	1	0.9803	97%
30	30	18	0.9615	1	0.9803	97%
35	35	21	0.9433	1	0.9708	96%
40	40	24	0.9433	1	0.9708	96%
45	45	27	0.9433	1	0.9708	96%
50	50	30	0.9433	1	0.9708	96%
60	60	36	0.909	1	0.9523	93%
70	70	42	0.909	1	0.9523	93%
80	80	48	0.8771	1	0.9345	91%
90	90	54	0.8771	1	0.9345	91%
100	100	60	0.8771	1	0.9345	91%

Skenario 1 menghasilkan nilai akhir yang telah diproses dari menghitung nilai *precision*, *recall*, *f-measure* dan akurasi. Pada tahap ini merupakan perhitungan menentukan n k-values awal kembali agar mendapatkan ketepatan sesuai kelas kategori. Pada proses menghitung *f-measure* mendapatkan hasil nilai akhir yang paling tinggi maupun yang paling rendah, dari hasil proses perhitungan *f-measure* yang mempunyai nilai tinggi dengan k-values awal 2, 4, 6, 8 dan 10 mendapatkan hasil nilai sebesar 1, dan dari hasil proses perhitungan *f-measure* yang mempunyai nilai rendah dengan k-values awal 80, 90 dan 100 mendapatkan hasil nilai sebesar 0.9345.

2. Hasil Pengujian Skenario 2

Tabel 3 Hasil Pengujian Skenario 2

k-values	n (k values baru)		Precision	Recall	F-Measure	Akurasi
	Isu SARA	Bukan Isu SARA				
2	2	1	1	1	1	100%
4	4	2	1	1	1	100%
6	6	4	1	1	1	100%
8	8	5	1	1	1	100%
10	10	6	0.9803	1	0.99	98%
15	15	9	0.9615	1	0.9803	97%

20	20	12	0.9615	1	0.9803	97%
25	25	15	0.9433	1	0.9708	96%
30	30	18	0.9259	1	0.9615	95%
35	35	21	0.909	1	0.9523	93%
40	40	24	0.909	1	0.9523	93%
45	45	27	0.909	1	0.9523	93%
50	50	30	0.909	1	0.9523	93%
60	60	36	0.909	1	0.9523	93%
70	70	42	0.8928	1	0.9433	92%
80	80	48	0.8771	1	0.9345	91%
90	90	54	0.862	1	0.9258	90%
100	100	60	0.862	1	0.9258	90%

Skenario 2 menghasilkan nilai akhir yang telah diproses dari menghitung nilai *precision*, *recall*, *f-measure* dan akurasi. Pada tahap ini merupakan perhitungan menentukan n k-values awal kembali agar mendapatkan ketepatan sesuai kelas kategori. Pada proses menghitung *f-measure* mendapatkan hasil nilai akhir yang paling tinggi maupun yang paling rendah, dari hasil proses perhitungan *f-measure* yang mempunyai nilai tinggi dengan k-values awal 2, 4, 6 dan 8 mendapatkan hasil nilai sebesar 1, dan dari hasil proses perhitungan *f-measure* yang mempunyai nilai rendah dengan k-values awal 90 dan 100 mendapatkan hasil nilai sebesar 0.9258.

3. Hasil Pengujian Skenario 3

Tabel 4 Hasil Pengujian Skenario 3

k-values	n (k values baru)		Precision	Recall	F-Measure	Akurasi
	Isu SARA	Bukan Isu SARA				
2	2	1	1	1	1	100%
4	4	2	1	1	1	100%
6	6	4	1	1	1	100%
8	8	5	1	1	1	100%
10	10	6	1	1	1	100%
15	15	9	1	1	1	100%
20	20	12	0.9803	1	0.99	98%
25	25	15	0.9803	1	0.99	98%
30	30	18	0.9803	1	0.99	98%
35	35	21	0.9803	1	0.99	98%
40	40	24	0.9615	1	0.9803	97%
45	45	27	0.9433	1	0.9708	96%
50	50	30	0.9433	1	0.9708	96%
60	60	36	0.9259	1	0.9615	95%
70	70	42	0.9433	1	0.9708	96%
80	80	48	0.9259	1	0.9615	95%
90	90	54	0.9259	1	0.9615	95%
100	100	60	0.909	1	0.9523	93%

Skenario 3 menghasilkan nilai akhir yang telah diproses dari menghitung nilai *precision*, *recall*, *f-measure* dan akurasi. Pada tahap ini merupakan perhitungan menentukan n k-values awal kembali agar mendapatkan ketepatan sesuai kelas kategori Pada proses menghitung *f-measure* mendapatkan hasil nilai akhir yang paling tinggi maupun yang paling rendah, dari hasil proses perhitungan *f-measure* yang mempunyai nilai tinggi dengan k-values awal 2, 4, 6, 8, 10 dan 15 mendapatkan hasil nilai sebesar 1, dan dari hasil proses perhitungan *f-measure* yang mempunyai nilai rendah dengan k-values awal 100 mendapatkan hasil nilai sebesar 0.9523.

4. Hasil Pengujian Skenario 4

Tabel 5 Hasil Pengujian Skenario 4

k-values	n (k values baru)		Precision	Recall	F-Measure	Akurasi
	Isu SARA	Bukan Isu SARA				
2	2	1	1	1	1	100%
4	4	2	1	1	1	100%
6	6	4	1	1	1	100%
8	8	5	1	1	1	100%
10	10	6	1	1	1	100%
15	15	9	1	1	1	100%
20	20	12	0.9803	1	0.99	98%
25	25	15	0.9803	1	0.99	98%
30	30	18	0.9615	1	0.9803	97%
35	35	21	0.9615	1	0.9803	97%
40	40	24	0.9615	1	0.9803	97%
45	45	27	0.9615	1	0.9803	97%
50	50	30	0.9615	1	0.9803	97%
60	60	36	0.9615	1	0.9803	97%
70	70	42	0.9615	1	0.9803	97%
80	80	48	0.9615	1	0.9803	97%
90	90	54	0.9615	1	0.9803	97%
100	100	60	0.9615	1	0.9803	97%

Skenario 4 menghasilkan nilai akhir yang telah diproses dari menghitung nilai *precision*, *recall*, *f-measure* dan akurasi. Pada tahap ini merupakan perhitungan menentukan n k-values awal kembali agar mendapatkan ketepatan sesuai kelas kategori. Pada proses menghitung *f-measure* mendapatkan hasil nilai akhir yang paling tinggi maupun yang paling rendah, dari hasil proses perhitungan *f-measure* yang mempunyai nilai tinggi dengan k-values awal 2, 4, 6, 8, 10 mendapatkan hasil nilai sebesar 1, dan dari hasil proses perhitungan *f-measure* yang mempunyai nilai rendah dengan k-values awal 30, 35, 40, 45, 50, 60, 60, 80, 90 dan 100 mendapatkan hasil nilai sebesar 0.9803.

5. Hasil Pengujian Skenario 5

Skenario 5 menghasilkan nilai akhir yang telah diproses dari menghitung nilai *precision*, *recall*, *f-measure* dan akurasi. Pada tahap ini merupakan perhitungan menentukan n k-values

awal kembali agar mendapatkan ketepatan sesuai kelas kategori. Pada proses menghitung *f-measure* mendapatkan hasil nilai akhir yang paling tinggi maupun yang paling rendah, dari hasil proses perhitungan *f-measure* yang mempunyai nilai tinggi dengan k-values awal 2, 4 dan 6 mendapatkan hasil nilai sebesar 1, dan dari hasil *f-measure* terendah dengan k-values awa dan dari hasil proses perhitungan *f-measure* yang mempunyai nilai rendah dengan k-values awal l 80, 90 dan 100 mendapatkan hasil nilai sebesar 0.9345.

Tabel 6 Hasil Pengujian Skenario 5

k-values	n (k values baru)		Precision	Recall	F-Measure	Akurasi
	Isu SARA	Bukan Isu SARA				
2	2	1	1	1	1	100%
4	4	2	1	1	1	100%
6	6	4	1	1	1	100%
8	8	5	0.9803	1	0.99	98%
10	10	6	0.9803	1	0.99	98%
15	15	9	0.9615	1	0.9803	97%
20	20	12	0.9803	1	0.99	98%
25	25	15	0.9803	1	0.99	98%
30	30	18	0.9803	1	0.99	98%
35	35	21	0.9803	1	0.99	98%
40	40	24	0.9615	1	0.9803	97%
45	45	27	0.9433	1	0.9708	96%
50	50	30	0.9433	1	0.9708	96%
60	60	36	0.9259	1	0.9615	95%
70	70	42	0.8928	1	0.9433	92%
80	80	48	0.8771	1	0.9345	91%
90	90	54	0.8771	1	0.9345	91%
100	100	60	0.8771	1	0.9345	91%

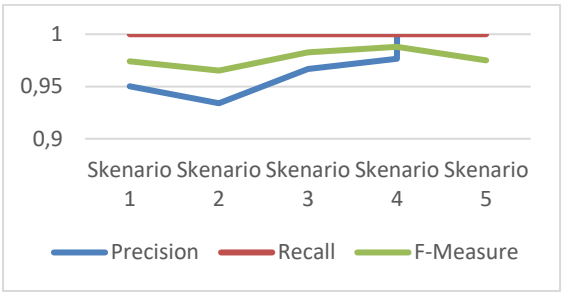
6. Tabel hasil akhir rata-rata proses perhitungan Precision, Recall dan F-Measure

Dibawah ini tabel 7 mendapatkan hasil akhiran rata-rata pengolahan uji *precision, recall* dan *f-measure* tiap skenario sudah diujikan:

Tabel 7 Rata-rata Precision, Recall dan F-Measure

Skenario	Precision	Recall	F-Measure	Akurasi
1	0.950338889	1	0.973994444	96%
2	0.933966667	1	0.965211111	95%
3	0.966628	1	0.98275	98%
4	0.976422	1	0.987944444	98%
5	17.1414	1	0.97502778	96%

Dari Tabel 7 yang mempunyai jumlah dokumen data latih 320 data maupun dokumen data uji 80 data pada pemrosesan awal menghitung *f-measure* menghasilkan akhiran skenario rata-rata yang terbaik berada skenario 4 yang menghasilkan rata-rata nilai sebesar 0.97502778.



Gambar 2 Grafik Rata-rata Hasil Precision, Recall, F-Measure

4.1 Analisis

Proses selanjutnya perhitungan pada klasifikasi algoritme Improved K-NN dan mendapatkan beberapa faktor yang mempengaruhi hasil pengujian. Hasil pengujian evaluasi data dokumen yang menggunakan 400 dokumen data yang diantaranya terdapat data latih 320 dokumen data maupun data uji 80 dokumen data, hingga hasil pengujian scenario 1 sampai dengan scenario 5 menghasilkan ketetapan hasil akhiran rerata *precision, recall, f-measure,* dan akurasi berbeda-beda pada tiap skenario.

Pada hasil akhir rata-rata pada masing-masing skenario yang nilai terendah pada proses pengujian *f-measure* menunjukkan pada skenario 2, dan *f-measure* tertinggi menunjukkan pada hasil skenario 4, pengujian setiap skenario diolah secara bergantian hingga tidak memiliki persamaan dokumen pada setiap skenarionya hingga 400 dokumen terpenuhi.

Proses uji pada tahapan *precision, recall, f-measure,* dan akurasi mendapatkan nilai n k-values pertama terkecil mengakibatkan hasil nilai menjadi paling rendah. Dari kejadian tersebut peneliti harus meneliti dengan baik agar dalam penentuan perhitungan n k-values awal harus teliti sehingga mendapatkan hasil kelas kategori yang sesuai.

5. KESIMPULAN

Perhitungan dalam metode algoritme *Improved KNN* bisa diterapkan ke dalam proses perhitungan klasifikasi isu sara menggunakan media sosial twitter. Dokumen teks tweet dihitung dan diolah melalui beberapa tahapan prose perhitungan preprocessing, pembobotaan term, dan nilai cosine similarity terhadap data dokumen latih. Selanjutnya dilakukan pengurutan pada kemiripan hasil token, dan menetukan n *k-values* awal untuk proses klasifikasi pada dokumen data.

Pada penelitian klasifikasi isu sara pada twitter berbahasa indonesia menggunakan metode *Improved K-NN*, proses klasifikasi didapatkan nilai rerata pada proses perhitungan *Precision* sebanyak 0.976422, *Recall* sebanyak 1, *F-Measure* sebanyak 0.987944444 dan Akurasi sebanyak 96%. Dari hasil tersebut setiap hasil dokumen data pada seluruh skenario berbeda-beda yang diacu pada pemilihan 400 dokumen data yang dilakukan pengujian hingga 400 dokumen tersebut terpenuhi semua.

6. DAFTAR PUSTAKA

Adikara, P. P., Perdana, R. S. & I., 2017. Model Vector Space. Malang: Fakultas Ilmu Komputer.

Bintang, 2013. Kominfo: Pengguna Internet di Indonesia 63 Juta Orang. [Online] Available at: <https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker> [Diakses 4 September 2018].

Feldman, R. & Sanger, J., 2006. Advanced Approaches in Analyzing Unstructured Data. [online] United States of America by Cambridge University Press, New Yor: Cambridge University Press. [Online] Available at: <http://www.cambridge.org/9780521836579>

Herdiawan, 2015. Analisis Sentimen Terhadap TELKOM Indihome berdasarkan Opini Publik menggunakan Metode Improved K-Nearest Neighbor. s.l.:s.n.

Herwijayanti, B., 2018. Klasifikasi berita online dengan menggunakan pembobotan tf-idf dan cosine similarity. s.l.:s.n.

Manning, C. D., Raghavan, P. & Schutze, H., 2009. An Introduction to Information Retrieval. England: Cambridge University Press..

Megantara, G., Kurniati, A. P. & Suryani, A. A., 2010. KLASIFIKASI TEKS DENGAN MENGGUNAKAN IMPROVED K-NEAREST NEIGHBOR ALGORITHM. s.l.:Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung.

Nathania, D. Z., 2017. Klasifikasi spam pada Twitter menggunakan Improved KNearest Neighbor. s.l.: Universitas Brawijaya.

Nur, M. Y. & Santika, D. D., 2011. Analisis Sentimen pada Dokumen berbahasa Indonesia dengan pendekatan Support Vector Machine. s.l.:s.n.

Puspitasari, A. A., t.thn. Klasifikasi Dokumen Tumbuhan Obat Menggunakan Metode Improved k-Nearest Neighbor. s.l.:Universitas Brawijaya.

Putri, P. A., t.thn. Implementasi Metode Improved K-Nearest Neighbor Pada Analisis Sentimen Twitter Berbahasa Indonesia. s.l.: Universitas Brawijaya.

Twitter, 2017. Getting started with Twitter. [Online] Available at: <https://support.twitter.com/articles/215585getting-started-with-twitter>[Diakses 3 Mei 2019].

XIong, F. & Liu, Y., 2014. Opinion formation on social media: an empirical approach. Chaos,. [Online] Available at: <http://www.ncbi.nlm.nih.gov/pubmed/24697392>.