Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 2, Februari 2019, hlm. 1235-1242 http://j-ptiik.ub.ac.id

Klasifikasi Kemacetan Lalu Lintas di Kota Malang Pada Sosial Media Twitter Menggunakan Metode *Improved K-Nearest Neighbor*

Riska Dewi Nurfarida[1], Indriati[2], Rizal Setya Perdana[3]

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

Abstrak

Twitter adalah jejaring sosial media yang mempunyai banyak pengguna yang dapat digunakan untuk media komunikasi. Dan dari twitter juga bisa didapatkan berbagai macam bentuk infomasi diantaranya yaitu opini yang bersifat negatif maupun positif dan berbagai macam informasi lainnya. Salah satu informasi yang bisa didaptkan dari Twitter yaitu informasi mengenai keadaan lalu lintas. Masyarakat Kota Malang menggunakan sosial media Twitter media untuk mendapatkan informasi tentang keadaan lalu lintas. Melalui akun @PuspitaFM, masyarakat Kota Malang saling berbagi informasi tentang keadaan lalu lintas di sekitar mereka. Dari akun @PuspitaFM tersebut setiap harinya akan berbagi *tweet* tentang keadaan lalu lintas di Kota Malang baik secara *tweet* langsung maupun tweet dari *follower* yang akan diretweet oleh akun @PuspitaFM. Dari seluruh *tweet* yang ada, terkadang terdapat kerancuan yang terjadi dalam pengkategorian macet atau tidak macet dalam tweet tersebut. Oleh karena itu, dilakukan pengklasifikasian *tweet* macet atau tidak macet sebagai solusi dari masalah tersebut. Terdapat beberapa proses yang dilakukan dalam penelitian ini, yaitu dimulai dari *prepocessing text* yang terbagi menjadi proses *cleansing, case folding,* tokenisasi, *filtering* dan *stemming*. Dari proses akan dilanjutkan dengan proses *term weighting* atau pembobotan, dilanjutkan dengan normalisasi, *cosine similiarity* dan proses klasifikasi dengan metode *Improved K-NN*. Hasil yang didapatkan dari penelitian ini yaitu nilai *recall* sebesar 0.42857, nilai *precision* sebesar 0.71428, nilai *f-measure* sebesar 0.53571 dan hasil akurasi terbaik sebesar 65.33%. Adapun data latih yang digunakan yaitu 600 dokumen *tweet*, dan data uji 150 dokumen *tweet.*

Kata kunci: *Information Retrieval,* Kemacetan Lalu Lintas, Twitter, *Improved K-Neares Neighbor*

Abstract

Twitter is a social media network that has many users that can be used for communication media. And from Twitter you can also get various forms of information including negative and positive opinions and various other types of information. One of the information that can be obtained from Twitter is information about traffic conditions. Malang City community uses Twitter social media as one of the media to get information about traffic conditions. Through the @PuspitaFM account, the people of Malang City share information about the state of traffic around them. From the @PuspitaFM account, every day I will share tweets about traffic conditions in Malang City either by tweeting directly or tweets from followers that will be retweeted by the @PuspitaFM account. Of all the tweets that exist, sometimes there is confusion that occurs in the categorization of traffic jams or not jammed in the tweet. Therefore, the classification of tweets is jammed or not jammed as a solution to the problem. There are several processes carried out in this study, namely starting from prepocessing text which is divided into cleansing, case folding, tokenisation, filtering and stemming processes. The process will continue with the term weighting or weighting process, followed by normalization, cosine similiarity and classification processes with the Improved K-NN method. The results obtained from this study are recall value of 0.42857, precision value of 0.71428, f-measure value of 0.53571 and the best accuracy of 65.33%. The training data used is 600 tweet documents, and 150 test data tweet documents.

Keywords: *Information Retrieval, Traffic congestion, Twitter, Improved K-Neares Neighbor*

1. PENDAHULUAN

Kemacetan lalu lintas adalah permasalahan yang terjadi karena jumlah kendaraan melampaui kapasitas jalan yang ada sehingga laju kendaraan menjadi terhambat. Permasalahan ini terjadi di kota-

kota besar, salah satunya yaitu Kota Malang. Adapun Kota Malang merpakan salah satu kota termacet yang ada di Indonesia setelah Kota Jakarta dan Kota Bandung.(Ramadhiani, 2018). Dengan perkembangan teknologi yang ada saat ini informasi tentang kemcetan bisa didapatkan melalui berbagai cara, salah satunya yaitu melalui media Twitter.

Twitter adalah jaringan sosial media yang mempunyai banyak pengguna yang dapat digunakan untuk media komunikasi. Dan dari twitter juga bisa didapatkan berbagai macam bentuk infomasi diantaranya yaitu opini yang bersifat negatif maupun positif dan berbagai macam informasi lainnya. Salah satu informasi yang bisa didaptkan dari Twitter yaitu informasi mengenai keadaan lalu lintas. Masyarakat Kota Malang menggunakan sosial media Twitter sebagai salah satu media untuk mendapatkan informasi tentang keadaan lalu lintas. Melalui akun @PuspitaFM, masyarakat Kota Malang saling berbagi informasi tentang keadaan lalu lintas di sekitar mereka. Dari akun @PuspitaFM tersebut setiap harinya akan berbagi *tweet* tentang keadaan lalu lintas di Kota Malang baik secara *tweet* langsung maupun tweet dari *follower* yang akan diretweet oleh akun @PuspitaFM. Beragamnya jenis tweet yang terdapat pada akun tersebut terkadang memiliki jenis tweet yang ambigu dalam menentukan kategori macet atau tidak macet, sehingga diperlukan adanya suatu penelitian untuk mengklasifikasikan tweet dalam menentukan kategori macet atau tidak macet.

Terdapat berbagai macam metode dalam melakukan pengklasifikasian teks, adapun metode tersebut terbagi menjadi 3 kelompok yakni klasifikasi teks dengan dasar statistic (Naïve Bayes, KNN CCV, SVM, dan sebagainya), kemudian klasifikasi teks berdasarkan koneksi (Artifical Neural Network), serta klasifikasi teks dengan dasar rule based (Decision Tree). Yang Yiming dan Xin Liu (1999) pada penelitiannya menyatakan metode klasifikasi berbasis statistik terutama KNN dan SVM terbukti mempunyai kinerja yang lebih baik dibandingkan metode lainnya (Ridok dan Latifah, 2015).

Metode yang digunakan dalam penelitian ini yaitu metode *Improved K-NN*. Metode tersebut adalah modifikasi dari metode *K-Nearest Neighbor* (K-NN). Metode K-NN memiliki kekurangan yaitu kurangnya ketepatan dalam menentukan kelas-kelas dari data kandidat hasil. Dari modifikasi metode K-NN tersebut (metode *Improved K-Nearest Neigbor)* dirasa dapat sebagai solusi yang ada pada metode K-NN (Megantara dkk, 2010). Perbedaan dari metode K-NN dan *Improved K-NN* yaitu terdapat pada penentuan k-value, pada metode K-NN k-value pada setiap kategori mempunyai *value* yang sama, sedangkan dalam metode *Improved K-NN* k-value masing-masing kategori mempunyai *value* yang berbeda dengan menyesuaikan banyaknya jumlah data latih (Puspitasari dkk, 2017). Dengan adanya perubahan k-value pada masing-masing kategori tersebut menghasilkan nilai akurasi yang lebih maksimal. Metode *Improved K-NN* dirasa merupakan metode yang sesuai untuk melakukan pengklasifikasian yang dapat menghasilkan kelas-kelas yang sesuai.

Pada penelitian yang dilakukan oleh Nathania dkk. (2018) mengenai klasifikasi spam pada twitter dengan menggunakan metode *Improved K-NN* dan metode K-NN sebagai metode pembanding. Penelitian ini dilakukan dengan jumlah data latih yaitu 500 dokumen *tweet.* Adapun hasil dari penelitian ini didapatkan hasill akurasi metode *Improved K-NN* sebbesar 92% dan metode K-NN sebesar 88%. Hal ini menunjukkan bahwa metode *Improved K-NN* mempunyai hasil yang lebih baik daripada metode K-NN.

Berdasarkan penelitian-penelitian tersebut , penulis menyarankan pendapat untuk melakukan penelitian klasifikasi kemacetan lalu lintas di Kota Malang pada sosial media twitter dengan menggunakan metode *Improved K-NN*.

DASAR TEORI

2. DASAR TEORI

2.1 *Information Retrieval*

Infromation Retrieval adalah suatu system untuk melakukan perankingan dokumen berdasarkan ketepatan terhadap *query* yang didapatkan dari pengguna. Hasil perangkingan yaitu dokumen yang mempunyai relevansi terhadap query. Dari

pengguna didapatkan tingkat relevansi yang bersifat subjektif yang dipengaruhi beberpa faktor, yaitu pewaktuan,sumber informasi, topik dan tujuan pengguna. (DwijaWisnu, 2015).

2.2 *Prepocessing Text*

Prepocesing Text adalah tahapan untuk memperoleh data yang terstruktur untuk memudahkan proses perhitungan selanjutnya (Nathania, 2018). Pada tahapn ini terdapat eberapa proses yaitu, *cleansing*, *case folding*, *filtering*, *tokenisasi* dan *stemming*.

1. *Cleansing*

   Cleansing adalah tahapan untuk membersihkan *noise* pada data yang dapat berupa *hashtag (#xxx)*, URL, angka, *username (@xxx)*, dan tanda baca.
2. *Case Folding*

   Merupakan tahapan untuk mengubah data dalam bentuk huruf kapital (*upper case*) menjadi bentuk huruf kecil (*lower case*).
3. Tokenisasi

   Tokenisasi adalah tahapan pemotongan kata pada dokumen dengan menggunaan spasi.
4. *Filtering*

   Filtering adalah tahapan untuk menghilangkan kata yang termasuk dalam datar stoplist. Kata yang masuk dalam daftar *stoplist* biasanya adalah kata yang tak bermakna dan diaggap tidak penting.
5. *Stemming*

Stemming adalah tahapan mengubah bentuk kata berimbuhan awal, tengah maupun akhir sehingga berubah menjadi kata dasar.

2.3 Pembobotan *(Term Weghting)*

Term Weighting adalah tahapan untuk mendapatkan nilai dari suatu *term* atau kata dengan cara pembobotan pada setiap *term* atau kata. Jumlah kemunculan *term* atau kata pada setiap dokumen disebut dengan *term frequency (TF)*. Nilai TF akan semakin bertambah apabila jumlah kemunculan *term* pada dokumen semakin tinggi. Proses untuk menghitung bobot jumlah kemunculan *term* atau kata pada seiap dokumen (Wtf) ditunjukkan pada persamaan (1).

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

$$1 + f(t,d) = \begin{cases} 1 + \log_{10} tf_{t,d}\ , & if\ tf_{t,d} > 0 \\ 0, & otherwise \end{cases}$$

Dimana:

- $W_{tf_{t,d}}$: Hasil dari pembobotan $tf_{t,d}$
- f(d,t) : frekuensi kemunculan *term t* pada dokumen *d*.
- $tf_{t,d}$: Frekuensi kemunculan *t* pada dokumen *d*

Adapun jumlah kemunculan *term* atau kata dalam sekumpulan dokumen disebut dengan *IDF* (*Inverse Document Frequency*) proses untuk menghitung kemunculan *term* atau kata pada sekumpulan dokumne ditunjukkan pada persamaan (2) (Fauzin, Arifin, Yuniarti, 2014).

$$IDF(t) = 1 + \log\ (\frac{Nd}{df(t)}) \quad (2)$$

Dimana:

- Nd: jumlah seluruh dokumen
- df(t): jumlah dokumen yang memiliki *term* t.

Untuk menggabungkan bobot padasetiap *term* pada masing-masing dokumen dilakuan dengan mengalikan bobot nilai TF dengan bobot niali IDF. Proses ini disebut juga dengan pembobotan TF-IDF yang persamaannya ditunjukkan pada persamaan (3) (Fauzin, Arifin, Yuniarti, 2014.

$$w(t,d) = W_{t,f}(t,d) \times idf_t \quad (3)$$

Dimana:

- $w(t,d)$: pembobotan TF-IDF
- $w_{tf}(t,d)$: pembobotan tft,d
- $idft$: invers nilai dft

Proses setelah menghitung pemobobptan TF-IDF adalah proses normalisasi yang persamaannya ditunjukkan pada persamaan (4) (Fauzin, Arifin, Yuniarti, 2014.

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

Proses yang dilakukan setelah proses normmalisasi yaitu proses untuk mengukur tingkat kemiripan antara dokumen dengan *query* yang disebut sebagai proses *cosine similiarity* yang dihitung dengan menggunakan persamaan (5)(Fauzi, Arifin, Yuniarti, 2014).

$$Cosine\ (d_i, q_i) = \quad (5)$$

$$\frac{q_i,d_i}{|q_i||d_i|} = \frac{\sum_{j=i}^{t}(q_{ij},d_{ij})}{\sqrt{\sum_{j=i}^{t}(q_{ij})^2.\sum_{j=i}^{t}(d_{ij})^2.}}$$

(2.7)

Dimana:

- q_{ij} = Bobot *j* pada dokumen i
- d_{ij} = Bobot *j* pada dokumen i

2.4 *Improved K-NN*

Metode *Improved K-NN* adalah modifikasi dari metode K-NN dimana k-value pada masing-masing kategori mempunyai nilai yang berbeda. Adapun perbedaan metode *Improved K-NN* dengan metode K-NN yaitu k-value pada metode KNN digunakan untuk seluruh kategori yang ada. sedangkan metode *Improved K-NN* k-value pada masing-masing kategori memiliki nilai yang berbeda (Baoli, Shiwen dan Qin, 2003). Adapun persamaan untuk mneghitung k-value baru pada masing-masing kateori yaitu ditunjukkan pada persamaan (6) (Herdiawan, 2015).

$$n = [\frac{k*N(C_m)}{maks\{N(C_m)|\ j=1.....N_c\}}] \quad (6)$$

Dimana:

- n : k-values baru
- k : k-values awal
- (Cm) : jumlah data latih kategori m
- maks N cm j=1……Nc} : jumlah data latih terbanyak pada seluruh kategori.

Setelah didapatkan k-value baru untuk masing-masing kategori, maka dilanjutkan dengan melakukan perbandingan nilai similiaritas pada setiap kategori. Proses tersebut dilakuakan untuk proses penentuan kategori pada data uji. Pada persamaan (7) menunjukkan proses perhitungan nilai probabiliras masing-masing kategori (Baoli, Shiwen dan Qin, 2003).

$$p(x, c_m) = \quad (7)$$

$$argmax_m \frac{\sum d_j \in top\ n\ KNN(c_m) sim(x,d_j) y(d_j.c_m)}{\sum d_j \in top\ n\ KNN(c_m) sim(x,d_j)}$$

Dimana:

- p(x,cm) : probabilitas data X menjadi anggota kategori cm
- (x,j) : kemiripan antara data X dengan data latih dj
- top n kNN : top n tetangga
- y(dj,cm) : fungsi atribut yang memenuhi kategori tertentu, akan bernilai 1 apabila data latih dj masuk anggota cm, dan bernilai 0 apabila data latih dh tidak masuk anggota cm.

Setelah menghitung nilai probabilitas pada masing-masing kategori, kemudian dilanjutkan dengan membandingkan hasil nilai prbabilitas pada setiap kategori. Hasil kategori yang diperoleh mengacu pada hasil nilai probabilitas yerbesar (Putri, 2013).

2.5 Pecision, Recall, F-Measure & Akurasi

Confusion matrix adalah tabel yang berfungsi sebagai perbandingan kategori aktual dengan kategori hasil prediksi (Nathania, Indriarti dan Bachtiar, 2018).

Tabel 1. *Confusion Matrix*

		Hasil Aktual	
		M	TM
Hasil Prediksi	M	TP	FP
	TM	FN	TN

Pada Tabel 1 terdapat nilai TP, FP, FN dan TN. Adapun M menunjukkan kategori macet dan TM menunjukkan kategori tidak macet. TP (*True Positive)* merupakan banyakya data pada hasil prediksi masuk kategori macet dan data pada hasil aktual masuk kategori macet. FP (*False Positive)* merupakan banyakya data pada hasil prediksi masuk kategori macet dan data pada hasil aktual masuk kategori tidak macet. FN (*False Negative)* merupakan banyakya data pada hasil prediksi masuk kategori tidak macet dan data pada hasil aktual masuk kategori macet. TN (*True Negative)* merupakan banyakya data pada hasil prediksi masuk kategori tidak macet dan data pada hasil aktual masuk kategori tidak macet.

Precision adalah tingkat keakuratan untuk mengetahui hasil kategori data yang diklasifikasikan sesuai dengan kategori sebenarnya. Persamaan *precission* ditunjukkan pada persamaan (8). *Recall* adalah parameter untuk mengetahui tingkat keberhasilan sistem untuk mengenali sebuah kategori. Persamaan recall ditunjukkan pada persamaan (9). F-*measure* merupakan gambaran pengaruh antara *precision* dan *recall* (Puspitasari, Santoso, Indriarti, 2018). Persamaan *f-measure* ditunjukkan pada persamaan (10).

$$precision = \frac{TP}{TP+FP} \quad (8)$$

$$recall = \frac{TP}{TP+FN} \quad (9)$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \quad (10)$$

Pada persamaan (11) ditunjukkan persamaan untuk mneghitung akurasi.

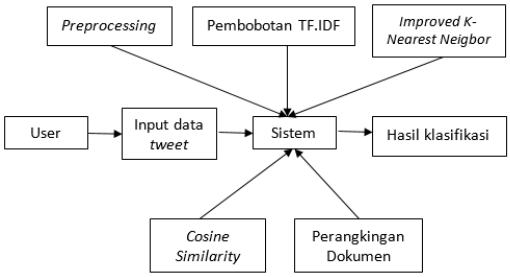
$$akurasi = \frac{TP+TN}{TP+FP+TN+FN} * 100\% \quad (11)$$

3. METODOLOGI

Pada bagian ini ditunjukkan rancangan penelitian, teknik pengumpulan data dan partisipan penelitian yang digunakan dalam penelitina ini.

3.1. Rancangan Penelitian

Rancangan peneltian merupakan gambaran mengenai cara kerja sitem yang digunakan dalam penelitian ini. Pada Gambar 1 dijelaskan bahwa *user* akan menginputka data *tweet* yang kemudian akan diproses melalui berbagai tahapan untuk mendaptkan hasil klasifikasi.



Gambar 1 Gambaran Sistem

3.2. Partisipan Penelitian

Partisipan penelitian dalam peneltian ini yaitu semua aktor yang terlibat dalam akun @PuspitaFM baik admin maupun follower yang mentweet tentang keadaan lalu lintas di Kota Malang. Kemudian terdapat 3 mahasiswa Teknik Informatika FILKOM Universitas Brawijaya yang terlibat dalam pengisian kuisioner data latih.

3.3. Teknik Pengumpulan Data

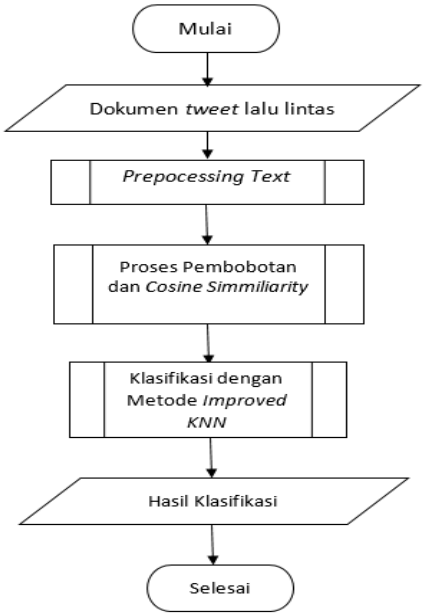
Data dari penelitian ini didapatkan dari akun twitter @PuspitaFM, dimana didapatkan 600 data tweet yang terdiri dari 400 data tidak macet dan 200 data macet.

4. PERANCANGAN

Pada bagian ini dijelaskan langkah-langkah apa saja yang dilakukan untuk membangun sistem ini.

4.1 Diagram Alir Sistem

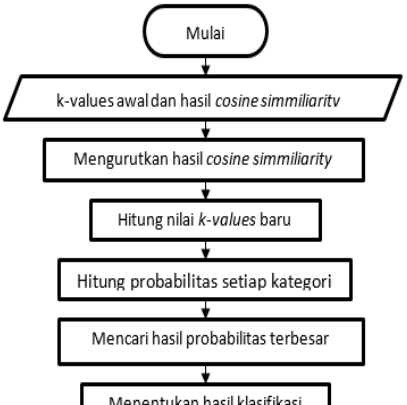
Pada bagian ni akan menjelaskan proses sistem yang dilihat secara keseluruhan. Proses ini dimulai dari dibutuhkan dokumen *tweet* lalu lintas yang kemudian akan diproses dengan *prepocessing text*, dilanjutkan dengan proses pembobotan dan *Cosine Similiarity* dengan menggunakan persamaan 5. Dlanjutkan dengan proses klasifikasi dengan metode *Improved K-NN* dengan menentukan *k-value* baru yang didapat dari persamaan 6 dilanjutkan perhitungan probabilitas masing-masing kategori denga persamaan 7.



Gambar 2 Diagram Alir Sistem Secara Keseluruhan

4.2 Diagram Alir Metode *Improved K-NN*

Metode *Improved K-NN* adalah modifikasi dari metode K-NN dimana k-value pada masing-masing kategori mempunyai nilai yang berbeda. Adapun perbedaan metode *Improved K-NN* dengan metode K-NN yaitu k-value pada metode KNN digunakan untuk seluruh kategori yang ada. sedangkan metode *Improved K-NN* k-value pada masing-masing kategori memiliki nilai yang berbeda (Baoli, Shiwen dan Qin, 2003). Adapun alur metode *Improved K-NN* ditunjukkan pada Gambar 3.



Gambar 3 Diagram Alir Metode *Improved K-NN*

5. PENGUJIAN DAN ANALISIS

Terdapat 7 skenario pengujina yang dilakukan dalam penelitian ini dengan tujuan mngetahui pengaruh dari jumlah data latih yang digunakan dan pengaruh k-value yang digunkan dalam penelitian ini terhadap 150 data uji yang ada. Pada Tabel 2 ditunjukkan skenario yang ada penelitian ini. Adapun k-value awal yang digunakan yaitu 2, 4, 6, 8, 10, 15, 20, 25, 30, 40, 45, 50, 75 dan 100

Tabel 2 Skenario Pengujian

Skenario	Data Latih			Data Uji		
	M	TM	Jumlah	M	TM	Jumlah
1	100	200	300	70	80	150
2	115	250	365	70	80	150
3	125	350	475	70	80	150
4	150	400	550	70	80	150
5	185	415	600	70	80	150
6	200	200	400	70	80	150
7	200	100	300	70	80	150

5.1. Perbandingan Metode *Improved K-NN* dan *K-Nearest Neighbor*

Dari ketujuh skenario pengujian yang telah dilakukan didaptkan hasil akurasi terbaik yaitu padaskenario 5. Pada skeanrio 5 digunakan 60data latih dengan jumlah data latih macet sebanyak 185 dan data latih tidak macet sebanyak 415. Skeanrio 5 ii akan dijadikan pemanding dengan metode K-NN. Adapun tabel perbandingan skenario 5 dengan metode *Improved K-NN* dan metode K-NN ditunjukkan pada Gambar 4.

k-values	Improved K-NN				K-NN			
	Precision	*Recall*	*F-Measure*	Akurasi	*Precision*	*Recall*	*F-Measure*	Akurasi
2	0,68085	0,45714	0,54701	64,67%	0,68085	0,45714	0,54701	64,67%
4	0,71429	0,42857	0,53571	65,33%	0,70000	0,40000	0,50909	64,00%
6	0,71429	0,42857	0,53571	65,33%	0,73684	0,40000	0,51852	65,33%
8	0,70000	0,40000	0,50909	64,00%	0,69048	0,41429	0,51786	64,00%
10	0,71429	0,42857	0,53571	65,33%	0,66667	0,37143	0,47706	62,00%
15	0,70588	0,34286	0,46154	62,67%	0,76667	0,32857	0,46000	64,00%
20	0,68571	0,34286	0,45714	62,00%	0,72727	0,22857	0,34783	60,00%
25	0,74194	0,32857	0,45545	63,33%	0,66667	0,17143	0,27273	57,33%
30	0,69565	0,22857	0,34409	59,33%	0,68750	0,15714	0,25581	57,33%
40	0,61111	0,15714	0,25000	56,00%	0,77778	0,10000	0,17722	56,67%
45	0,68750	0,15714	0,25581	57,33%	0,77778	0,10000	0,17722	56,67%
50	0,71429	0,14286	0,23810	57,33%	0,80000	0,11429	0,20000	57,33%
75	0,80000	0,11429	0,20000	57,33%	0,90000	0,12857	0,22500	58,67%
100	0,85714	0,08571	0,15584	56,67%	0,85714	0,08571	0,15584	56,67%
Rata-rata	0,71592	0,28878	0,39151	61,19%	0,74540	0,24694	0,34580	60,33%

Gambar 4 Tabel Perbandingan Metode Improved K-NN dan K-Nearest Neighbor

Dari tabel perbandingan tersebut dapat diketahui bahwa dengan metode *Improved K-NN* didapatkan rata-rata *f-measure* lebih baik yaitu dengan nilai 0,39151, sedangkan metode K-NN memperoleh nilai *f-measure* yaitu 0,34580. Rata-rata hasil akurasi dengan metode *Improved K-NN* juga menunjukkan nilai yang labih baik dengan nilai 61,19%, sedangkan dengan metode K-NN didaptkan nilai yaitu 60,33%. Hal tersebut menunjukkan bahwa metode *Improved K-NN* menghasilka nilai yang lebih baik daripada metode K-NN.

5.2. Analisis

Pada pengujian yang telah dilakukan didapatkan pada skenario 1, 2, 3, 4 dan 5 memiliki nilai f-measure tertinggi pada nilai k-awal bernilai 2. Skenario ini memiliki jumlah perbandingan data tidak macet lebih besar daripada data macet. Pada skenario 6 nilai f-measure tertinggi ada pada nilai k-awal bernilai 15. Skenario 6 memiliki jumlah perbandingan data tidak macet dan data macet berimbang. Pada skenario 7 memiliki nilai f-measure tertinggi ada pada nilai k-awal bernilai 40. Skenario 7 memiliki jumlah perbandingan data macet lebih besar daripada data macet. Maka, dapat disimpulkan bahwa perbandingan jumlah data yang digunakan mempengaruhi nilai f-measure pada nilai k-awal. Hal ini dapat dilihat pada skenario 1, 2, 3, 4 dan 5 yang memiliki jumlah perbandingan data tidak macet lebih besar daripada data macet memiliki nilai f-measure tertinggi pada nilai k-awal kecil yaitu 2, sedangkan pada skenario 6 dan 7 memiliki jumlah perbandingan data tidak macet dan data macet berimbang dan memiliki jumlah perbandingan data macet lebih besar daripada data macet memiliki nilai f-measure tertinggi

pada nilai k-awal bernilai 15 dan 40. Oleh karena itu, penentuan jumlah data latih harus dilakukan dengan teliti karena perbedaan dan perbandingan jumlah data latih dapat mempengaruhi hasil pengujian.

Pada data latih dan data uji terdapat tweet yang mengandung keterangan waktu berupa jam. Pada proses prepocessing keterangan berupa jam tersebut tidak dihilangkan karena pada tweet tersebut hanya untuk menunjukkan pola kapan terjadinya macet atau tidak macet. Hal tersebut tidak mempengaruhi hasil dari klasifikasi karena pada dasarnya tweet tersebut sama saja dengan tweet yang tidak mengandung keterangan waktu berupa jam.

Pada penelitian ini memiliki hasil akurasi dan selisih dengan metode KNN terbilang rendah, dimana hasil akurasi tertinggi hanya bernilai 61,19%. Hal ini disebabkan karena adanya term atau kata yang muncul pada kedua kategori. Misalnya, term yang menunjukkan nama jalan muncul di kategori macet dan tidak macet. Hasil klasifikasi dari metode Improved K-NN dan KNN memuliki selisih rata-rata yang terbilang rendah hanya bernilai 0,86%. Hal ini disebabkan karena adanya nilai k-value baru yang sama dengan nilai k-value awal yang ditetapkan. Dan nilai k-value baru tersebut terdapat pada ketegori dengan jumlah data yang lebih banyak.

6. KESIMPULAN

Metode *Improved K-NN* bisa digunakan dalam melakuakn pengklasifikasian kemacetan lalu lintas dengan data yang berupa *tweet*. Data *tweet* ini sebelumnya dilakukan beberapa proses untuk mendapatkan hasil klasifikasinya. Proses tersebut yaitu *prepocessing text*, pembobotan (*term weighting)*, normalisasi,*cosine similiarity*, mengurutkan tingkat kemiripan, menentukan *k-value* baru dan menghitung probabilitas masing-masing kategori. Adapun dari hasil pengujian didapatkan hasil terbaik dengan nilai akurasi 65.33%, nilai *f-measure* 0.53714, nilai *recall* 0.428571 dan nilai *precision* 0.714286. Berdasarkan perbandingan hasil pengujian skenario terbaik dengan *Improved K-NN* didapatkan rata-rata hasil akurasi yang lebih baik daripada metode K-NN. Hal ini menunjukkan bahwa *Improved K-NN* lebih baik daripada K-NN, namun dengan nilai selisih yang tidak begitu besar.

7. DAFTAR PUSTAKA

Baoli, L., Shiwen, Y. dan Qin, L., 2003. An Improved k-Nearest Neighbor Algorithm for Text Categorization. [online] Tersedia di: <https://pdfs.semanticscholar.org/490a/b325ba480f6fb71cdbb5f87ff4cb70918686.pdf> [Diakses 25 Januari 2018]

Claudy, Y.I., Perdana, R.S., Fauzi, M.A., 2018. *Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN)*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 8, Agustus 2018, hlm. 2761-2765.

Fauzi, M.A., Arifin, A.Z., Yuniarti, A., 2014. *Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab*. LONTAR KOMPUTER VOL. 5, NO. 2, AGUSTUS 2014.

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

Nathania, D.Z., Indriarti., Bachtiar, F.A., 2018. *Klasifikasi Spam Pada Twitter Menggunakan Metode Improved KNN*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 10, Oktober 2018, hlm. 3948-3956.

Puspitasari, A.A., Santoso Edy., Indriati., 2018. *Klasifikasi Dokumen Tumbuhan Obat Menggunakan Metode Improved k-Nearest Neighbor*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol. 2, No. 2, Februari 2018, hlm. 486-492.

Putri, P.A., Ridok Achmad., Indriati., 2017. *IMPLEMENTASI METODE IMPROVED K-NEAREST NEIGHBOR PADA ANALISIS SENTIMEN TWITTER BERBAHASA INDONESIA*. [online] Tersedia di: < https://www.academia.edu/11472083/IMPLEMENTASI_METODE_IMPROVED_K-NEAREST_NEIGHBOR_PADA_ANALISIS_SENTIMEN_TWITTER_BERBAHASA_INDONESIA?auto=downlo

ad> [Diakses 25 Januari 2018]

Ridok, Ahmad., Latifah, Ritnani., 2015. *Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN*. Konferensi Nasional Sistem & Informatika 2015.

Rodiansyah, S.F., Winarko ,E., 2013. *Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification*. IJCSS Vol. 7, No.1, Januari 2013, hlm 13-22.

Sremanthy, J. Balamurugan, P.S., 2012. *AN EFFICIENT TEXT CLASSIFICATION USING KNN AND NAIVE BAYESIAN*. International Journal on Computer Science and Engineering (IJCSE). Coimbatore, India.