REPOSITOR, Vol. 2, No. 2, Februari 2020, Pp. 185-192
ISSN : 2714-7975
E-ISSN : 2716-1382

Analisa Sentimen Tweet Berbahasa Indonesia dengan Menggunakan Metode Pembobotan Hybrid TF-IDF pada Topik Transportasi Online

Sari Wahyunita[1], Yufis Azhar[2], Nur Hayatin[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
wahyunitas@gmail.com[1], yufis@umm.ac.id[2], noorhayatin@gmail.com

Abstrak

Beberapa tahun belakangan ini, muncul perusahaan-perusahaan penyedia jasa transportasi yang menggunakan aplikasi berbasis android dalam proses pelayanannya atau biasa disebut dengan transportasi online. Hal ini dilakukan untuk meningkatkan pelayanan terhadap pengguna jasa transportasi. Hadirnya transportasi online seperti Gojek, Grab dan Uber menimbulkan masalah sosial antara supir transportasi online dan supir transportasi non aplikasi. Penyebabnya dikarenakan sebagian besar masyarakat beralih menggunakan transportasi online, sehingga pendapatan supir transportasi non aplikasi menurun. Pada penelitian ini, dilakukan analisa sentimen terhadap tweet berbahasa Indonesia tentang transportasi online dengan menggunakan metode pembobotan Hybrid TF-IDF dan kNN sebagai metode klasifikasinya. Hasil terbaik dari pengujian cross validation pada uji variable k adalah k=5 dengan nilai akurasi 70%, presisi kelas positif 68%, presisi kelas negatif 75%, recall kelas positif 82%, recall kelas negatif 59%, f-measur kelas positif 74% dan f-measure kelas negatif 65%.

Kata Kunci: *Hybrid TF-IDF, Analisa Sentimen, Transportasi Online, kNN, Cross Validation*

Abstract

In recent years, there are same new companies which uses android applications in their service process or commonly called with online based transportation. This case for improve their service to online based transportations customers. The presence of online based transportation like Gojek, Grab, and Uber inflict social problem between online based transportations drivers with non-aplication based transportations drivers. The is due to most of the people change over to online based transportation, so non-aplication based transportations driver income decreased. In this research, sentiment analysis against Indonesian tweets about online based transportation by using Hybrid TF-IDF weighting method and kNN classification method. The best cross validation result on the test k variable are k=5 with accuracy value 70%, positive class precision 68%, negative class precision 75%, positive class recall 82%, negative class recall 59%, positive class f-measure 74%, and negatif class f-measure 65%.

Keywords: *Hybrid TF-IDF, Sentiment Analysis, Online based Transportation, kNN, Cross Validation*

1. Pendahuluan

Salah satu kebutuhan pokok masyarakat Indonesia adalah transportasi. Bagi sebagian besar masyarakat transportasi sangat penting untuk menunjang kegiatan sehari-hari mereka. Namun, tidak semua masyarakat memiliki tranportasi pribadi. Sehingga mereka harus menggunakan transportasi umum, seperti ojek, angkot, taksi komersial dan lain-lainnya.

Beberapa tahun belakangan ini, muncul perusahaan-perusahaan penyedia jasa transportasi yang menggunakan aplikasi berbasis android dalam proses pelayanannya atau biasa disebut dengan transportasi *online*. Hal ini dilakukan untuk meningkatkan pelayanan terhadap pengguna jasa transportasi. Ada 3 perusahaan besar yang menyediakan jasa transportasi *online* di Indonesia, yaitu GO-JEK, Grab dan Uber. Hadirnya transportasi *online* di Indonesia disambut dengan antusias oleh manyarakat Indonesia.

Besarnya antusias masyarakat terhadap transportasi online menimbulkan masalah sosial antara supir transportasi online dan supir transportasi non aplikasi. Supir transportasi non aplikasi seperti ojek, angkot dan taksi komersial merasa lahan pekerjaan mereka diambil oleh para supir transportasi online. Hal ini berdampak pada turunnya pendapatan supir transportasi non aplikasi

perhari. Penurunan pendapatan supir transportasi non aplikasi menyentuh angka 20% bahkan lebih [1].

Melihat permasalah yang timbul karena hadirnya transportasi online menimbulkan pertanyaan, seberapa besar respon masyarakat terhadap pelayanan transportasi online tersebut sehingga trasportasi online dapat menggeser kepopularitasan transportasi non aplikasi. Respon tersebut tidak jarang diungkapkan melalui media sosial, salah satunya adalah Twitter. Twitter merupakan salah satu media sosial yang memungkinkan penggunanya untuk menuliskan pesan ke dalam 140 karakter. Twitter di dirikan oleh Jack Dorsey, dan resmi di luncurkan pada tanggal 21 Maret 2006 yang ditandai dengan *tweet* (kicauan) pertama dari Jack Dorsey [2].

Berdasarkan data statistika bulan Mei 2016 oleh website statistika dot com, Indonesia menduduki peringkat ke-3 untuk pengguna aktif media sosial Twitter [3]. Hal ini membuktikan bahwa media sosial Twitter di Indonesia cukup popular dan tidak menuntut kemungkinan melalui Twitter kita dapat mengetahui respon masyarakat terhadap suatu permasalahan yang sedang terjadi di Indonesia.

Banyak penelitian di bidang data mining yang melakukan analisa sentimen pada Twitter. Hal ini dikarenakan dalam penelitian, Twitter merupakan sebuah indikator yang baik [4]. Salah satu jurnal yang membahas tentang penelitian analisa sentimen Twitter mengenai penyedia jasa transportasi adalah "Penerapan Algoritma Genetika untuk Seleksi Fitur Pada Analisa Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naïve Bayes" yang dilakukan oleh Risa Wati. Dimana penelitian tersebut bertujuan untuk mengetahui kualitas layanan jasa maskapai penerbangan dari *tweet* dengan menggabungkan metode Genetika dan Naïve Bayes pada proses klasifikasi [5].

Ada 2 cara untuk melakukan analisa sentimen terhadap teks, yaitu dengan menggunakan metode *lexicon based* atau metode pembobotan. Salah satu metode pembobotan yang dapat digunakan adalah Hybrid TF-IDF. Metode pembobotan Hybrid TF-IDF telah digunakan pada penelitian mengenai "Peringkasan Sentimen Ekstrasif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity" yang dilakukan oleh David Haryalesmana Wahid dan Azhari SN pada studi kasus artis Agnes Monica [6].

Usulan penelitian ini bertujuan untuk melakukan analisa sentimen pada data *tweet* berbehasa Indonesia dengan topik transportasi online dengan menggunakan metode pembobotan Hybrid TF-IDF. *Output* dari penelitian ini adalah berupa pengelompokan *tweet* berdasar negatif dan positif.

2. Metode Penelitian

Pada bagian ini akan menjelaskan tahapan yang dilakukan dalam analisa sentimen *tweet* mengenai transportasi *online* dengan menggunakan metode Hybrid TF-IDF.

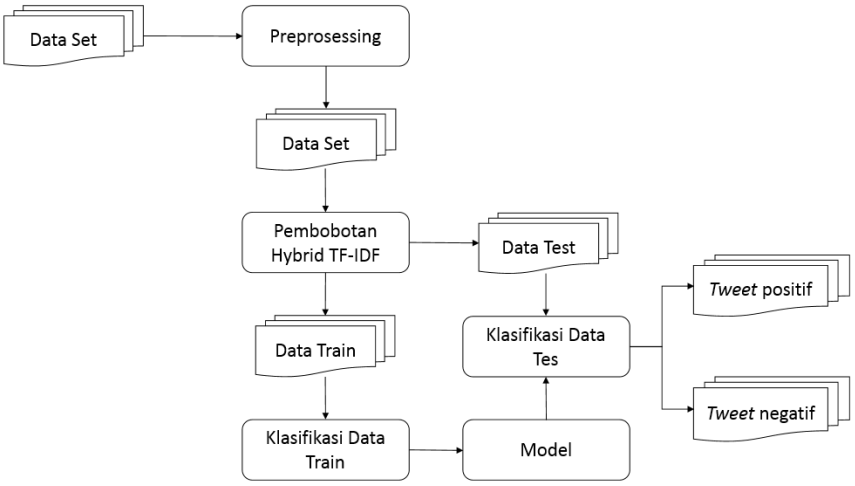
2.1. *Crawling* Data dan Analisa Data

Penelitian ini menggunakan data yang diambil dari Twitter dengan kriteria *tweet* mengenai transportasi online dan berbahasa Indonesia. *Crawling* dilakukan dengan menggunakan bahasa pemograman Python dan menggunakan *library* Tweepy. *Crawling* adalah suatu proses untuk mengambil data pada suatu database. Untuk mengakses database Twitter diperlukan *key* yang didapatkan dengan mendaftarkan proyek kita pada http://apps.twitter.com/. Kata kunci yang digunakan untuk *crawling* data *tweet* mengenai transportasi *online* adalah "gojekindonesia", "grabID", dan "uber_IDN". Hasil *crawling.*

Data yang digunakan sebagai data set dalam melakukan analisa sentimen ini berupa data *tweet* berbahasa Indonesia yang didapat dari media sosial Twitter. *Tweet* yang digunakan adalah *tweet* mengenai transportasi *online* sebanyak 700 *tweet.* Dimana dari 700 dataset tersebut dibagi menjadi 2 bagian, yaitu 600 *tweet* untuk data *train* dan 100 *tweet* untuk data *test.*

2.2. Rancangan Proses Klasifikasi

Rancangan proses klasifikasi adalah gambaran urutan tahap yang dilakukan sistem untuk melakukan klasifikasi dengan menggunakan metode kNN yang ditunjukan pada Gambar 1. Sistem dikembangan dengan menggunakan bahasa pomograman Python.



Gambar 1. Worfkflow Proses Analisa dan Perancangan

2.3. *Preprocessing*

Preprosessing adalah tahap awal untuk menyetarakan data yang telah dikumpulkan, hal ini juga berjutuan untung menghilangkan noise [7]. Tahapan yang dilakukan pada proses *preprocessing* disesuaikan dengan kebutuhan penelitian yang dilakukan. Mengacu pada jurnal yang berjudul "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia dengan Maximum Entropy dan Support Vector Machine", tahapan di dalam *preprocessing* pada data set berupa text diantaranya adalah sebagai berikut:

a. *Case folding*
   Pada tahap ini merupakan tahap paling awal, dimana tahap ini dilakukan untuk menyamakan besar kecilnya huruf pada teks yang akan diolah. Hal ini bertujuan untuk menyelaraskan bentuk huruf pada semua teks.
b. *Cleansing*
   Pada tahap ini dilakukan penghapusan tanda baca yang tidak penting, seperti koma (,), titik (.), dan lainnya.
c. *Tokenizing*
   Pada tahap ini merupakan tahapan dimana semua kalimat dipecah menjadi kata perkata sehingga dapat dilakukan pembobotan pada setiap kata yang terkandung dalam kalimat.
d. *Filtering*
   Pada tahap ini merupakan penghapusan URL, *mention* dan *hashtag*. Seperti " http://site.com ", "@uber_IDN", dan "#gojek".
e. *Replacement* atau *Stemming*
   Pada tahap dilakukan untuk menormalisasi kata-kata yang tidak baku menjadi kata yang baku, hal ini juga bertujuan untuk memudahkan pada proses klasifikasi. Seperti "melihat" menjadi "lihat". Dalam penelitian ini, *stemming* dilakukan dengan menggunakan *library* Sastrawi *stemmer*. *Library* Sastrawi *stemmer* menerapkan algoritma yang berbasis Nazief dan Adriani, kemudian ditingkatkan oleh Algoritma CS (Confix Stripping), kemudian ditingkatkan lagi oleh algortima ECS (Enhanced Confix Stripping), lalu ditingkatkan lagi oleh Modified ECS.

2.4. Pembobotan Hybrid TF IDF

Hybrid TF-IDF adalah metode pembobotan yang dikembangkan dari metode TF-IDF. Pada dasarnya TF-IDF merupakan teknik pembobotan yang berbasis statistik, dimana setiap kalimat akan diberikan bobot lalu kalimat tersebut akan diurutkan berdasarkan bobotnya. Untuk memcari bobot dengan metode TF-IDF dapat dilihat pada Persamaan 1 [8].

$$TF\_IDF = tf_{ij} * \log_2 \frac{N}{df_j} \quad (1)$$

Dimana tf_{ij} adalah jumlah *term* pada 1 data atau dokumen, *N* adalah jumlah data atau dokemen dan df_j adalah jumlah data/kaliam yang mengandung *term*.

Namun, metode TF-IDF kurang baik jika digunakan untuk pembobotan pada *tweet* karena jumlah kata pada *tweet* tidak lebih dari 140 karakter, sehingga 5*tweet* hanya mengandung beberapa kata saja. Oleh karena itu, Sharifi *et al* mengusulkan algoritma *Hybrid TF-IDF* [8]. Pada metode pembobotan *Hybrid TF-IDF* setiap *tweet* dianggap sebagai dokumen terpisah, namun perhitungan frekuensi *term* dilakukan pada keseluruhan *tweet*. Sehingga nilai TF menjadi normal dan tidak kehilangan property IDF [8] [9]. Berdasarkan jurnal yang berjudul "*Expriment in Microblog Summarization"* yang dijulis oleh Sharifi *et al* persamaan untuk perhitungan bobot term bias dilihat pada Persamaan 2, Persamaan 3, dan Persamaan 4 [8].

$$W(w_i) = tf(w) \cdot \log_2(idf(w_i)) \quad (2)$$

$$tf(w) = \frac{Jumlah\ kemunculan\ term\ di\ semua\ tweet}{jumlah\ term\ di\ semua\ tweet} \quad (3)$$

$$idf(w_i) = \frac{jumlah\ tweet}{jumlah\ tweet\ yang\ memuat\ term} \quad (4)$$

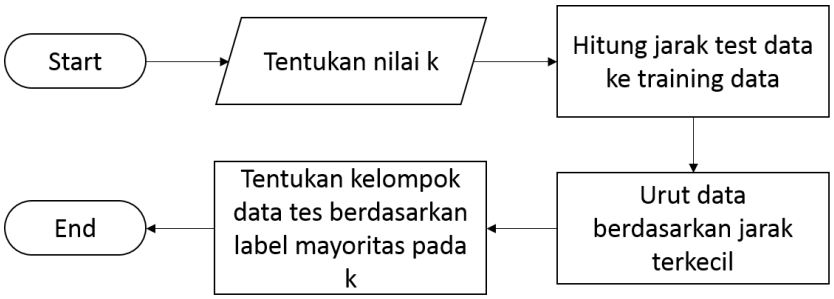
2.5. Klasifikasi kNN (*k-Nearest Neighbor*)

k-Nearest Neighbor (kNN) merupakan salah satu metode yang menggunakan algoritma *supervised.* Dimana hasil *query instance* akan diklasifikasikan berdasarkan *class* mayoritas [10]. Prinsip kerja kNN adalah mencari jarak yang terdekan antara data yang akan dievaluasi kengan *k* tetangga terdekatnya. Tahapan dalam melakukan klasifikasi dengan menggunakan metode kNN adalah sebegai berikut:

a. Pertama tentukan parameter nilai *k.* Dimana nilai *k* merupakan jumlah tetangga paling dekat.
b. Setelah itu, menghitung *euclidean distance* objek terhadap data *training* yang telah ditentukan sebelumnya. Menurut M Ilyas Sikki untuk menghitung *euclidean distance* dapat digunakan Persamaan 5 [10].

$$D(a,b) = \sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \quad (5)$$

c. Selanjutnya, hasil dari perhitungan *188uclidean distance* diurutkan secara *ascending* atau dari nilai paling kecil ke paling besar.
d. Setelah diurutkan, data yang belum terklasifikasi dapat diprediksi kategorinyanya berdasarkan kategori *nearest neighbor* yang paling mayoritas.



Gambar 2. Flowchat Algoritma KNN

Algoritma kNN pada Gambar 2 memiliki kelebihan, yaitu tangguh terhadap data *training* yang memiliki banyak noise dan efektif bila memproses data *training* yang besar. Sedangkan kelemahan algoritma kNN adalah peru ditentukannya nilai *k, training* berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil terbaik. Selain itu, biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap *query instance* pada keseluruhan data *training* [10].

2.6. *Cosine Similarity*

Selain menggunakan *euclidean distance* untuk menghitung jarak, metode yang dapat digunakan adalah *cosine similarity*. Metode *cosine similarity* merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antara dua buah objek. Secara umun metode ini didasarkan pada *vector space cosine similarity measure*. Metode *cosine similarity* ini menghitung

similarity antara dua buah objek (misalkan D1 dan D2) yang dinyatakan dalam dua buah *vector* dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebegai ukuran [11]. Persamaan *cosine similarity* dapat dilihat pada Persamaan 6.

$$S(D_i, Q_i) = \frac{\sum(D_i \times Q_i)}{\sqrt{(\sum D_i^2 \times \sum Q_i^2)}} \quad (6)$$

2.7. Pengujian

Pada penelitian ini, pengujian dilakukan secara intrinsik. Pengujian intrinsik dilakukan dengan cara menghitung akurasi dan *f-measure* yang di dapat dari perhitungan *precision* dan *recall. Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh system. *Recall* adalah tingkat keberhasilan system dalam menemukan kembali sebuah informasi. Untuk menghitung akurasi, digunakan Persamaan 7 sebagai berikut.

$$Akurasi = \frac{jumlah\ tweet\ positif + negatif\ yang\ dipisah\ dengan\ benar}{jumlah\ total\ tweet} \quad (7)$$

Sedangkan untuk menghitung nilai F-measure digunkan persamaan 10, dengan presisi dan *recall* menggunakan Persamaan 8 dan Persamaan 9 sebagai berikut.

$$Presisi = \frac{jumlah\ tweet\ positif\ atau\ negatif\ yang\ dipisah\ dengan\ benar}{jumlah\ tweet\ yang\ dipisahkan\ positif\ atau\ negatif} \quad (8)$$

$$recall = \frac{jumlah\ tweet\ positif\ atau\ negatif\ yang\ dipisah\ dengan\ benar}{jumlah\ tweet\ positif\ atau\ negatif\ yang\ sebenarnya} \quad (9)$$

$$F - measure = 2 \times \frac{presisi \times recall}{(presisi + recall)} \quad (10)$$

Selain itu untuk mendapatkan nilai akurasi, presisi, *recall,* dan *f-measure* yang akurat, maka dilakukan pengujian dengan metode *cross validation.* Metode *cross validation* adalah metode pengujian yang dilakukan untuk menprediksi *error* pada data *test* [11]. Dalam pengujian ini, data set akan dibagi menjadi k. Dari data yang telah dibagi, k-1 bagian akan menjadi data *train* dan 1 bagian akan menjadi data train. Hal ini disebut dengan k-fold.

Selanjutnya pada data set yang telah dibagi dengan k dilakukan proses penyilangan dimana data *train* dijadikan sebagai data *test* begitu juga data *test* menjadi data *train*, proses ini dilakukan sebanyak k kali. Setelah itu, hitung rata-rata nilai akurasi, presisi, *recall,* serta *f-measure* pada masing-masing data *test* yang telah disilang dan hasilnya dirata-rata.

3. Hasil Penelitian dan Pembahasan

Dari hasil klasifikasi data set dengan menggunakan metode pembobotan Hybrid TF IDF, maka selanjutnya dilakukan pengujian untuk menghitung akurasi, presisi, *recall,* dan *f-measure* dengan menggunkan metode *cross validation*. Dimana pembaginya (*k-fold*) adalah *k-fold*=7, hal ini karena data set berjumlah 700 *tweet*. Sehingga masing-masing jumlah data di setiap pembaginya adalah 100 *tweet*.

Selain itu, pengujian juga dilakukan dengan nilai k pada klasifikasi kNN yang berbeda-beda. Nilai k yang digunakan ada 3, yaitu 5, 7 dan 9. Hasil dari pengujian adalah Tabel 1.

Pada Tabel 1 hasil dari pengujian menunjukkan bahwa Data 5 dapat dikatakan memiliki nilai penguji paling baik dari semua aspek pengujian kecuali nilai *recall negative* yang mana nilai paling tinggi ada pada Data 6. Sedangkan data yang memiliki nilai pengujian paling tidak baik adalah Data 3.

Pada Tabel 2 hasil dari pengujian menunjukkan bahwa Data 6 dapat dikatakan memiliki nilai penguji paling baik dari semua aspek pengujian kecuali nilai presisi negative dan *recall* positif yang mana nilai paling tinggi ada pada Data 5. Sedangkan data yang memiliki nilai pengujian paling tidak baik adalah Data 3.

Table 1. Hasil Pengujian Dengan Metode Cross Validation K-Fold=7 Dengan K=5

Data	Akurasi	Presisi Positif	Presisi Negatif	Recall Positif	Recall Negatif	F-measure Positif	F-measure negatif
Data 1	0.67	0.63	0.73	0.80	0.54	0.71	0.62
Data 2	0.69	0.64	0.79	0.86	0.52	0.74	0.63
Data 3	0.51	0.51	0.52	0.78	0.24	0.61	0.33
Data 4	0.74	0.69	0.83	0.88	0.60	0.77	0.70
Data 5	0.81	0.77	0.86	0.88	0.74	0.82	0.80
Data 6	0.77	0.76	0.78	0.78	0.76	0.77	0.77
Data 7	0.74	0.73	0.75	0.76	0.72	0.75	0.73
Rata - Rata	0.70	0.68	0.75	0.82	0.59	0.74	0.65
Persentase	70%	68%	75%	82%	59%	74%	65%

Table 2. Hasil Pengujian Dengan Metode Cross Validation K-Fold=7 Dengan K=7

Data	Akurasi	Presisi Positif	Presisi Negatif	Recall Positif	Recall Negatif	F-measure Positif	F-measure negatif
Data 1	0.66	0.63	0.70	0.76	0.56	0.69	0.62
Data 2	0.66	0.63	0.70	0.76	0.56	0.69	0.62
Data 3	0.50	0.50	0.50	0.76	0.24	0.60	0.32
Data 4	0.73	0.69	0.79	0.84	0.62	0.76	0.70
Data 5	0.81	0.78	0.84	0.86	0.76	0.82	0.80
Data 6	0.82	0.81	0.83	0.84	0.80	0.82	0.82
Data 7	0.70	0.69	0.71	0.72	0.68	0.71	0.69
Rata - Rata	0.70	0.68	0.73	0.79	0.60	0.73	0.65
Persentase	70%	68%	73%	79%	60%	73%	65%

Table 3. Hasil Pengujian Dengan Metode Cross Validation K-Fold=7 Dengan K=9

Data	Akurasi	Presisi Positif	Presisi Negatif	Recall Positif	Recall Negatif	F-measure Positif	F-measure negatif
Data 1	0.60	0.59	0.61	0.66	0.54	0.62	0.57
Data 2	0.65	0.62	0.70	0.78	0.52	0.69	0.60
Data 3	0.48	0.49	0.46	0.70	0.26	0.57	0.33
Data 4	0.72	0.68	0.78	0.82	0.62	0.75	0.69
Data 5	0.84	0.79	0.90	0.92	0.76	0.85	0.83
Data 6	0.82	0.81	0.83	0.84	0.80	0.82	0.82
Data 7	0.77	0.80	0.75	0.72	0.82	0.76	0.78
Rata - Rata	0.70	0.68	0.72	0.78	0.62	0.72	0.66
Persentase	70%	68%	72%	78%	62%	72%	66%

Pada Tabel 3 hasil dari pengujian menunjukkan bahwa Data 5 dapat dikatakan memiliki nilai penguji paling baik dari semua aspek pengujian kecuali nilai presisi positif dan *recall* negatif. Sedangkan data yang memiliki nilai pengujian paling tidak baik adalah Data 3.

Pada Tabel 4 hasil dari pengujian menunjukkan bahwa Data 6 dapat dikatakan memiliki nilai penguji paling baik dari semua aspek pengujian. Sedangkan data yang memiliki nilai pengujian paling tidak baik adalah Data 3.

Sedangkan untuk nilai rata-rata pengujian *cross validation* dengan nilaik k yang berbeda-beda ditunjukan pada Tabel 5. Dari Tabel 5 dapat diketahui bahwa nilai rata-rata akurasi dan rata-rata presisi positif memiliki nilai yang sama untuk masing-masing hasil pengujian pada nilai k yang berbeda-meda. Namun jika dilihat dari nilai rata-rata pengujian yang lainnya, maka untuk

klasifikasi yang menggunakan nilai k=5 memiliki hasil pengujian yang paling baik. Sedangkan hasil pengujian yang paling tidak baik adalah klasifikasi yang menggunakan nilai k=15.

Table 4. Hasil Pengujian Dengan Metode Cross Validation K-Fold=7 Dengan K=15

Data	Akurasi	Presisi Positif	Presisi Negatif	Recall Positif	Recall Negatif	F-meansure Positif	F-meansure negatif
Data 1	0.60	0.60	0.60	0.60	0.60	0.60	0.60
Data 2	0.56	0.55	0.58	0.68	0.44	0.61	0.50
Data 3	0.45	0.46	0.41	0.66	0.24	0.55	0.30
Data 4	0.68	0.67	0.69	0.70	0.66	0.69	0.67
Data 5	0.80	0.83	0.78	0.76	0.84	0.79	0.81
Data 6	0.85	0.91	0.81	0.78	0.92	0.84	0.86
Data 7	0.74	0.77	0.71	0.68	0.80	0.72	0.75
Rata - Rata	0.67	0.68	0.65	0.69	0.64	0.68	0.64
Persentase	67%	68%	65%	69%	64%	68%	64%

Table 5. Hasil Pengujian Dengan Metode Pembobotan Hybrid T F IDF

Nilai K	k = 5	k = 7	k = 9	k=15
AVG Akurasi	70%	70%	70%	67%
AVG Presisi Positif	68%	68%	68%	68%
AVG Presisi Negatif	75%	73%	72%	65%
AVG Recall Positif	82%	79%	78%	69%
AVG Recall Negatif	59%	60%	62%	64%
AVG F-measure Positif	74%	73%	72%	68%
AVG F-measure Negatif	65%	65%	66%	64%

4. Kesimpulan

4.1. Kesimpulan

Berdasarkan hasil penelitian analisa sentimen terhadap transportasi *online* dengan menggunakan metode pembobotan Hybrid TF-IDF dapat disimpulkan bahwa

1. Klasifikasi Hasil terbaik dari pengujian *cross validation* pada uji variable k adalah k=5 dengan nilai akurasi 70%, presisi kelas positif 68%, presisi kelas negatif 75%, *recall* kelas positif 82%, *recall* kelas negatif 59%, *f-measur* kelas positif 74% dan *f-measure* kelas negatif 65%.

4.2. Saran

Saran yang dapat diberikan terhadap pengembangan analisa sentimen dengan menggunakan metode Hybrid TF-IDF adalah sebegai berikut:

1. Menggunakan data set yang lebih banyak dari data set yang digunakan pada penelitian ini.
2. Mengembangkan persamaan Hybrid TF-IDF agar hasil pengujiannya lebih baik.
3. Melakukan pengujian dengan nilai k yang lebih beragam.

Daftar Notasi

tf_{ij} : jumlah *term* pada 1 data atau dokumen.
N : jumlah data atau dokemen.
df_j : jumlah data/kaliam yang mengandung *term*.
$S(D_i, Q_i)$: *cosine similarity.*

Referensi

[1] M. Idris, "Pendapatan Operator Taksi Menurun Hingga 20% dengan Kehadiran Uber dan GrabCar," detiknews, 15 Maret 2016. [Online]. Available: http://news.detik.com/berita/3165540/pendapatan-operator-taksi-menurun-hingga-20-dengan-kehadiran-uber-dan-grabcar. [Accessed 9 Mei 2017].

[2] Twitter, "Pencapaian Twitter," 30 Juni 2016. [Online]. Available: https://about.twitter.com/id/company/press/milestones. [Accessed 11 Maret 2017].

[3] Statistika, "Statistika: Number of Active Twitter Users in Leading Markets as of May 2016 (in Millions)," Mei 2016. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed 12 Maret 2017].

[4] J. Weng, E.-P. Lim and J. Jiang, "TwitterRank: Finding Topic-sensitive Influential Twitters," *WSDM'10,* pp. 261-270, 2010.

[5] R. Wati, "Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naive Bayes," *Evolusi,* vol. 4, pp. 26-32, 2016.

[6] D. H. Wahid and A. SN, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *IJCCS,* vol. 10, pp. 207-218, 2016.

[7] M. I. Sikki, "Pengenalan Wajah Menggunakan K-Nearest Neighbour Dengan Praproses Transformasi Wavelet," *Jurnal Paradigma,* vol. X, pp. 159-172, 2009.

[8] G.Vinodhini and RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 6, pp. 282-292, 2012.

[9] B. Liu, Sentiment Analysis and Opinion Mining, San Rafael: Morgan&Claypool Publishers, 2012.

[10] O. Nurdiana, Jumadi and D. Nursantika, "Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Padaaplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia," *JOIN6,* vol. I, pp. 59-63, 2016.

[11] H. Leidiyana, "Penerapan Algoritma K-Nearest Neighbor untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor," *J Piksel,* vol. 1, pp. 65-76, 2013.