IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA

Muhammad Rivki, Adam Mukharil Bachtiar

Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Komputer Indonesia,
Jl. Dipati Ukur No. 112-114-116, 40132 Bandung, Indonesia

E-mail: muhammad.Rivki@outlook.com, adam@email.unikom.ac.id

Abstract

In this digital age, where marketing strategies are continuously growing, many entrepreneurs start to utilize social media as one of tools to do marketing strategy. Twitter is one of their preference of social media as a medium to sell products. Unfortunately, Twitter does not provide feature to facilitate users in doing promotions, such as providing information about followers are more likely being active in Twitter and then categorizing the interest of followers accordingly. In order to overcome this problem, a Twitter client that can classify users' followers to be able to facilitate the promotion on Twitter. One way to categorize consumer interest on Twitter is by using text mining. The K-Nearest Neighbor algorithm is one of algorithms that can be used to perform classification task. Development of Twikipedia system with K-Nearest Neighbor algorithm is able to classify users' followers and facilitate the former in doing promotion on Twitter.

Keywords: *K-Nearest Neighbor, Social Media Analytics, Text Mining, Twitter*

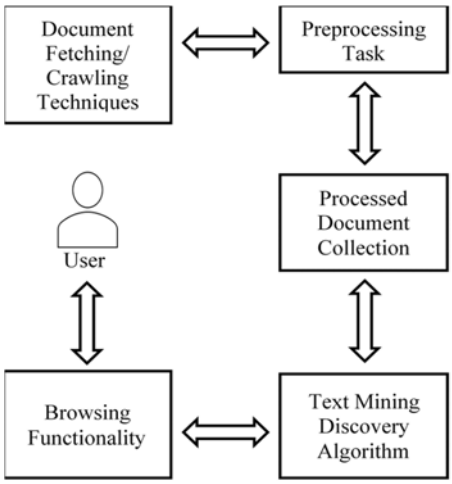
Abstrak

Di era digital ini, dimana strategi marketing terus berkembang, banyak pengusaha yang mulai memanfaatkan media sosial sebagai salah satu alat untuk melakukan strategi pemasaran. Twitter adalah salah satu media sosial yang digunakan sebagai media untuk memasarkan produk mereka. Sayangnya, Twitter tidak memberikan fitur untuk memudahkan penggunanya dalam melakukan promosi seperti memberikan informasi tentang waktu keaktifan *follower* serta mengkategorikan sesuai dengan ketertarikan dari *follower.* Untuk dapat mengatasi permasalahan tersebut, dibutuhkan sebuah Twitter *client* yang dapat melakukan klasifikasi terhadap *follower* dari pengguna dan memudahkan cara promosi di Twitter. Salah satu cara untuk mengkategorikan ketertarikan konsumen di Twitter adalah dengan menggunakan *text mining.* Algoritma K-Nearest Neighbor adalah salah satu algoritma yang bisa dimanfaatkan untuk implementasi pengklasifikasiannya. Pembangunan sistem Twikipedia dengan algoritma K-Nearest Neighbor mampu mengklasifikasikan *follower* dari pengguna dan memudahkan pengguna dalam melakukan promosi di Twitter.

Kata Kunci: *K-Nearest Neighbor, Social Media Analytics, Text Mining, Twitter*

1. Pendahuluan

Dalam dunia bisnis, *marketing* merupakan hal yang paling *crusial* atau penting. *Marketing* atau pemasaran adalah aktivitas, serangkaian institusi dan proses menciptakan, mengkomunikasikan, menyampaikan dan mempertukarkan tawaran (*offerings*) yang bernilai bagi pelanggan, klien, mitra, dan masyarakat umum [1]. Di era digital ini, dimana strategi marketing terus berkembang, banyak pengusaha yang mulai memanfaatkan media sosial sebagai salah satu alat untuk melakukan strategi pemasaran. Dari fakta tersebut timbul sebuah fenomena baru, yaitu *"Buzz Marketing"* atau *"Viral Marketing"* yang merupakan teknik pemasaran produk atau jasa untuk menghasilkan bisnis melalui informasi dari akun sosial media yang satu ke akun sosial media lainnya [2]. Salah satu strategi pemasaran atau *marketing (manual)* yang biasa digunakan adalah dengan cara promosi. Banyak cara yang bisa digunakan untuk melakukan promosi, salah satunya adalah dengan menyebarkan *flyer* maupun menyebarkan informasi dari mulut ke mulut. Namun dalam kenyataannya masih banyak pengusaha yang salah dalam melakukan promosi, seperti kesalahan dalam iklan yang terlalu mengawang dan terlalu mencakup terlalu banyak segmen (*general*) oleh karenanya iklan ini tidak sampai kepada konsumen. Selain dalam hal promosi, masih ada beberapa kesalahan yang biasa dilakukan yakni salah menentukan target pasar, salah dalam melihat perspektif pasar dan salah dalam melakukan promosi. Semua kesalahan tersebut menyebabkan banyak pengeluaran dan waktu yang terbuang



Gambar 1. Gambaran Metodologi Penelitian

sia-sia. Oleh karena itu dalam sebuah *marketing* diperlukan strategi yang tepat. Kesesuaian antara produk dan jasa yang dipromosikan dengan akun yang akan menjadi target promosi merupakan faktor yang sangat berpengaruh dalam melakukan promosi untuk meningkatkan keefektifan dan keefisiensian dari promosi tersebut.

Di samping itu sejak kemunculannya, media sosial selalu mendapatkan perhatian khusus dari masyarakat. Terbukti bahwa pengguna media sosial semakin bertambah dari waktu ke waktu. Berdasarkan data yang ada pada tahun 2013, pengguna media sosial di Indonesia telah mencapai 95% dari 63 juta pengguna internet dan akan terus bertambah [3]. Dari sekitar 32.000 sampel survei yang telah dilakukan, penggunaan sosial media mencapai dua kali lipat dari pada waktu yang digunakan untuk mengonsumsi media tradisional seperti televisi ataupun radio [4]. Hal tersebut telah membuat media sosial berkembang sangat pesat, karena di dalamnya terdapat berbagai hal mulai dari informasi, hiburan, edukasi, hobi, sampai hal pribadi. Sosial media yang kini telah menjadi rutinitas masyarakat, membuat kepopuleran televisi, radio, ataupun majalah menjadi berkurang. Salah satu sosial media yang paling berkembang pesat adalah *Twitter*. Menurut versi *Social Times*, *Twitter* termasuk ke dalam 3 besar sosial media dengan pertumbuhan yang paling cepat [5]. *Twitter* merupakan sebuah media sosial yang digunakan sebagai alat untuk berbagi berbagai macam informasi mengenai banyak hal yang menjadi perhatian pengguna. Dengan kepopulerannya tersebut banyak pemilik usaha yang mulai menggunakan *Twitter* sebagai media untuk promosi [6]. Salah satu alasan penggunaan *Twitter* sebagai media promosi adalah karena dengan menggunakan media sosial ini para pemilik usaha tidak perlu mengeluarkan dana yang besar untuk promosi.

Sayangnya, *twitter* tidak memberikan fitur untuk memudahkan penggunanya melakukan promosi seperti melakukan *tweet* yang terjadwal, memberikan informasi tentang waktu keaktifan *user*, kategorisasi *user* dan lain sebgaianya yang membantu memudahkan pengguna *twitter* untuk media promosi. Oleh karena itu untuk menghindari kesalahan yang sama dengan sistem *marketing* tradisional dan memudahkan pengguna *twitter* yang menggunakan *twitter* sebagai mediap promosi maka diperlukan sebuah sistem yang dapat digunakan untuk menentukan waktu aktif pengguna, menganalisis pengguna dan mengategorikan kebiasaan pengguna *Twitter*. Sistem tersebut akan menghasilkan *knowledge* yang akan dimanfaatkan sebagai pendukung media promosi agar iklan menjadi tepat sasaran yaitu hanya kepada calon konsumen yang sekiranya mempunyai peluang besar untuk menjadi konsumen dilihat dari segi *interest* dan kebutuhannya.

2. Metode Penelitian

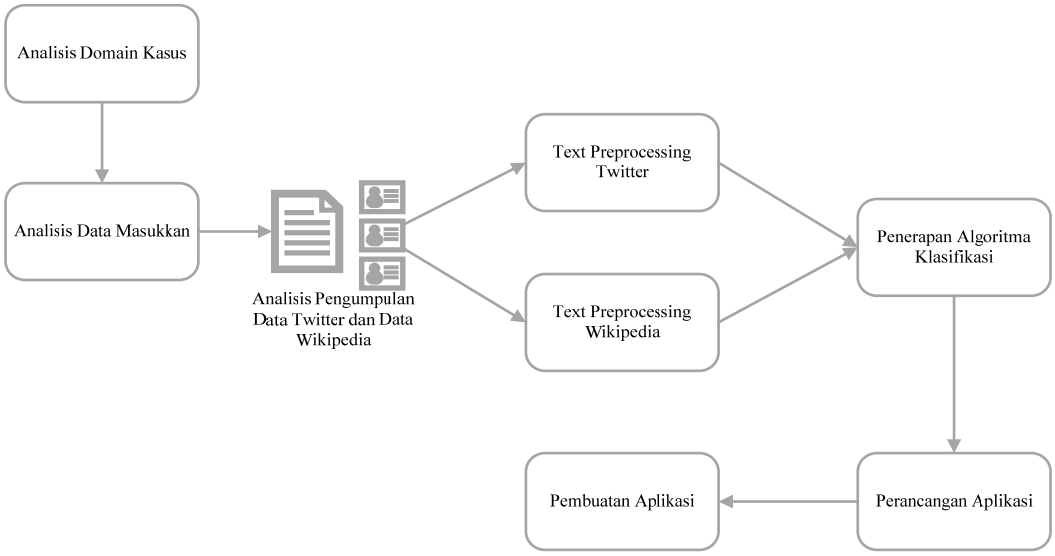
Metode penelitian yang digunakan adalah metode penelitian terapan deskriptif. Metode penelitian terapan bertujuan untuk menerapkan, menguji, dan mengevaluasi suatu teori dalam memecahkan suatu masalah. Langkah-langkah yang akan dilakukan pada penelitian ini mengacu pada [7]. Urutan dari metodologi penelitian yang dilakukan digambarkan pada Gambar 1. Kemudian, alur disesuaikan dari penelitian yang dilakukan sehingga menghasilkan alur seperti yang ditunjukkan Gambar 2.

Text Mining

Text mining yang juga dikenal dengan *text data mining* atau pencarian pengetahuan di basis data textual adalah sebuah proses untuk melakukan pencarian pengetahuan yang berfokus kepada data yang berbentuk dokumen atau teks, dengan tujuan untuk mengekstrak informasi yang berguna dan mengidentifikasinya. *Text mining* mempunyai kesamaan dengan *data mining*. Keduanya memiliki tujuan yang sama yaitu untuk memperoleh pengetahuan dan informasi dari sekumpulan data yang besar. Yang membedakan antara *text mining* dengan *data mining* adalah pada *data mining* masukan data terstruktur sedangkan *text mining* masukan datanya tidak terstruktur [8].

Text *Preprocessing*

Tahapan *text preprocessing* diperlukan untuk membersihkan sumber data dari data yang tidak diperlukan [8]. Proses ini bertujuan agar data yang digunakan nantinya bersih dari *noise* atau ciri-ciri yang tidak berpengaruh pada proses-proses selan-



Gambar 2. Gambaran Alur Penelitian

jutnya seperti *link*, "@", "RT", *stopword*. Proses *preprocessing* juga mempunyai tujuan agar data yang digunakan memiliki dimensi yang lebih kecil dan lebih terstruktur, sehingga dapat diolah lebih lanjut. Tahap *preprocessing* yang digunakan dapat dilihat pada Gambar 3.

Algoritma K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (KNN) adalah algoritma yang digunakan untuk melakukan klasifikasi terhadap suatu objek, berdasarkan *k* buah data latih yang jaraknya paling dekat dengan objek tersebut. Syarat nilai *k* adalah tidak boleh lebih besar dari jumlah data latih, dan nilai *k* harus ganjil dan lebih dari satu. Dekat atau jauhnya jarak data latih yang paling dekat dengan objek yang akan diklasifikasi dapat dihitung dengan menggunakan metode *cosine similiarity* [9].

Cosine similiarity merupakan salah satu cara atau metode yang dapat digunakan untuk melihat sejauh mana kemiripan isi antar dokumen. Dalam hal ini *cosine similiarity* berfungsi untuk menguji ukuran yang dapat digunakan sebagai interpretasi kedekatan jarak berdasarkan kemiripan dokumen. Persamaan(1) berikut ini adalah rumus untuk menghitung jarak pada algoritma KNN dengan metode *cosine similiarity*:

$$\cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n}(Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(D_i)^2}} \qquad (1)$$

Dimana,
$\cos(\theta_{QD})$ = Kemiripan *Q* terhadap dokumen *D*
Q = Data Uji
D = Data Latih
n = Banyaknya Data

Confussion Matrix

Confussion Matrix merupakan sebuah tabel yang terdiri atas banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi, tabel ini diperlukan untuk menentukan kinerja suatu model klasifikasi [10]. Metode ini sering digunakan dengan kasus *multiple classifiers* atau kelas yang lebih dari dua. Maka dari itu metode ini cocok untuk digunakan dalam penelitian ini untuk mengukur seberapa akurat hasil klasifikasi dari model yang telah dibuat. Perhitungan akurasi dengan menggunakan tabel *confussion matrix* adalah sebagaimana dirumuskan oleh persamaan (2).

$$Akurasi = \frac{\mathrm{F11} + \mathrm{F\ 00}}{\mathrm{F\ 11} + \mathrm{F\ 10} + \mathrm{F\ 01} + \mathrm{F\ 00}} \qquad (2)$$

3. Hasil dan Pembahasan

Analisis Masalah

Analisis masalah merupakan bentuk penjabaran mengenai masalah yang ada sebulm dibangunnya perangkat lunak *twikipedia* ini. Analisis masalah yang ada meliputi hal-hal sebagai berikut: 1) dari beberapa *twitter client* yang sudah ada, tidak adanya *twitter client* yang memiliki fitur untuk melakukan kategorisasi terhadap *follower* yang dimiliki penggunanya. Sehingga pengguna *twitter* yang menggunakan *twitter* sebagai media promosi akan merasa sulit untuk menentukan pengguna sesuai target promosi; 2) *Twitter client* yang ada tidak

TABEL 1
TABEL REST PUBLIK TWITTER

Metode	Path	Request/15 mnt/pengguna	Deskripsi
GET	application/rate_limit_status	180	Mengembalikan nilai limit dalam penggunaan API atau status penggunaan API pada setiap path.
GET	search/tweets	180	Mengembalikan tweets yang cocok dengan kata kunci yang diminta
GET	favorites/list	15	Mengembalikan 20 tweets disukai yang paling terbaru
GET	followers/ids	15	Mengembalikan follower dengan id
GET	followers/list	15	Mengembalikan data follower dari pengguna
GET	friends/list	15	Mengembalikan data following dari pengguna
GET	lists/list	15	Mengembalikan data list yang dimiliki oleh pengguna
GET	lists/members	180	Mengembalikan data member dari sebuah list
GET	lists/statuses	180	Mengembalikan data status dari timeline pengguna
GET	statuses/retweeters/ids	15	Menampilkan id pengguna yang me-retweet
GET	statuses/home_timeline	15	Mengembalikan data tweets yang paling terbaru di home pengguna
GET	statuses/retweets/:id	15	Mengembalikan id pengguna lain yang melakukan retweet
GET	statuses/show/:id	180	Mengembalikan tweets berdasarkan id lengkap dengan data pengguna
GET	statuses/pengguna_timeline	180	Mengembalikan data timeline dari pengguna
POST	statuses/update	-	Melakukan tweet
POST	statuses/retweet/:id	-	Melakukan retweet pada id tweet spesifik
POST	direct_messages/new	-	Melakukan direct message ke pengguna lain
GET	trends/available	15	Mengembalikan data trending topics

memiliki fitur untuk mengirim *tweet* atau mem-*broadcast* ke ba-nyak pengguna, sehingga pengguna *twitter* akan mengalami kesulitan untuk melakukan *tweet* ke semua pengguna yang sesuai dengan target promosi; 3) dibutuhkannya fitur untuk menentukan waktu yang tepat atau banyak user yang aktif untuk melakukan promosi di *twitter* karena jika tepat user yang membaca promosi akan banyak.

Berdasarkan masalah-masalah tersebut maka dibutuhkan sebuah perangkat lunak yang dapat melakukan kategorisasi terhadap pengguna *twitter* yang mem-*follow* pengguna dan dapat menentukan kapan waktu yang tepat untuk melakukan *tweet* promosi di *twitter.*

TABEL 2
DATA TRAINING

Tweets	Kategori	Fitur
T1	Travel	[naik, gunung, sumedang]
T2	Travel	[pertama, kali, naik, gunung, otot, tarik, waktu, tengah, tengah, hutan]
T3	Kuliner	[malam, tutup, nasi, bakar, ayam, taliwang, makan, santap, siomay, bakso, soto]
T4	Kuliner	[makan, malam, nasi, bakar, soto, ayam, enak, tahun, makan, soto]
T5	Olahraga	[bola, kantuk, kaya, kopi, cimahpar, endah, sukabumi]
T6	OlahRaga	[siluet, logo, nba, jerry, west, pemain, basket, legendaris, lama]

TABEL 3
DATA TESTING

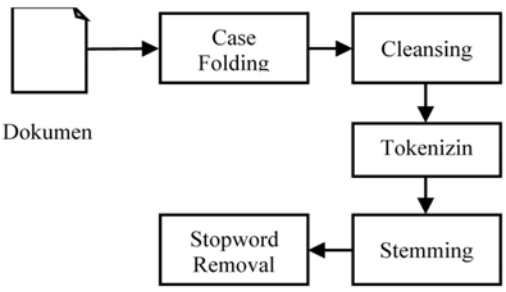
Data Testing	Kategori	Fitur
T	?	[lama, makan, nasi, soto, ayam, lamongan]

Analisis Arsitektur Sistem

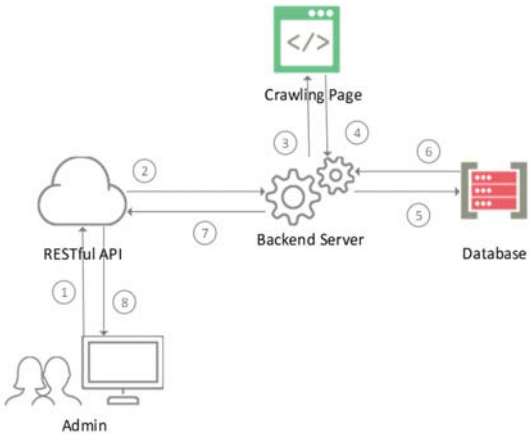
Tahap analisis arsitektur sistem adalah tahapan untuk mendapatkan gambaran umum system yang akan dibangun. Baik itu gambaran pengguna hingga transaksi data yang akan terjadi nantinya. Arsitektur sistem perangkat lunak yang akan dibangun dibagi menjadi dua yaitu dari sisi admin dan dari sisi pengguna akhir (Gambar 4 dan 5).

Analisis Pengumpulan Data Twitter

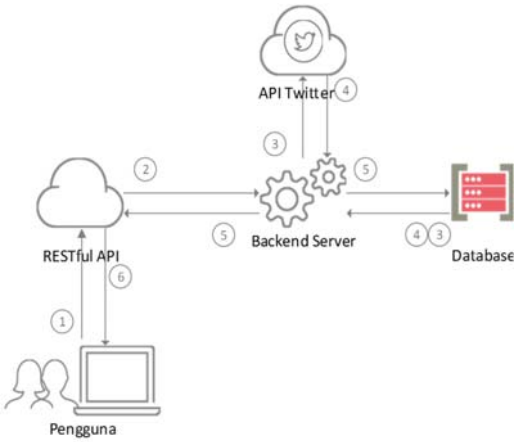
Data *twitter* dapat diperoleh dengan berbagai cara, salah satunya adalah dengan memanfaatkan RESTful APIs yang sudah disediakan oleh *twitter* dan dapat digunakan secara publik. *Twitter* menyediakan berbagai macam data dalam RESTful APIs mereka, seperti mendapatkan data *follower*, *follo-*



Gambar 3. Tahapan Preprocessing Tweet



Gambar 4. Arsitektur Sistem Admin



Gambar 5. Arsitektur Sistem Pengguna

wing, *user timeline* dan lain sebagainya namun dengan batasan harus memiliki *key* yang diberikan oleh *twitter* agar dapat mengakses RESTful APIs. Detail RESTful APIs yang disediakan oleh *twitter* dapat dilihat pada Tabel 1.

Analisis Text *Preprocessing*

Pada tahapan ini data yang sudah didapatkan dari Twitter akan melalui proses *preprocessing* untuk diberishkan datanya dari data yang sebelumnya tidak terstruktur menjadi data yang lebih ter-struktur agar memudahkan saat proses klasifikasi. Tahapan *prerprocessing* pada data yang didapat dari Twitter yaitu *case-folding, cleansing, tokenizing, stemming* lalu *stopword-removal*. Gambar 3 menunjukkan gambaran mengenai alur *preprocessing* data yang bersumber dari *twitter*.

Berikut adalah penjelasan masing-masing tahapan:

Case Folding: *case folding* adalah proses mengubah huruf kapital dari data masukkan menjadi huruf kecil semua (*lowercase*). Hal ini dilakukan agar semua huruf pada data masukkan menjadi seragam.

Cleansing: pada tahapan ini semua karakter non alfabet akan dihapus dengan tujuan mengurangi *noise* sehingga karakter khusus, URL, email, angka dan simbol akan dihapus.

Tokenizing: tokenizing merupakan proses untuk memisahkan kata yang dipisahkan oleh spasi. Hal ini dilakukan agar tahap *preprocessing* selanjutnya dapat berjalan.

Stemming: stemming merupakan proses untuk menghilangkan kata yang berimbuhan menjadi bentuk kata dasarnya.

Stopword Removal: pada tahap ini *tweets* masih mengandung kata yang dianggap tidak memberikan pengaruh dalam menentukan klasifikasi. Kata-kata tersebut dimasukkan ke dalam daftar *stopword*. Jika termasuk ke dalam *stopword* maka kata tersebut akan dihapus atau dihilangkan dari *tweet.*

Analisis Algoritma Klasifikasi

Pada tahap ini sebelum melakukan klasifikasi dengan algoritma KNN diperlukan pembobotan pada data *training*. Hal tersebut diperlukan karena algoritma KNN hanya dapat melakukan klasifikasi pada data yang berupa angka maka perlu adanya tahapan untuk mengubah data menjadi angka. Metode TF-IDF adalah metode yang digunakan untuk mendapatkan bobot dari data *training* yang akan menentukan klasifikasi pada data *testing* nantinya. Tabel 2 dan 3 menunjukkan data *training* dan *testing* yang dihasilkan metode TF-IDF dan dilakukan pada proses *preprocessing*.

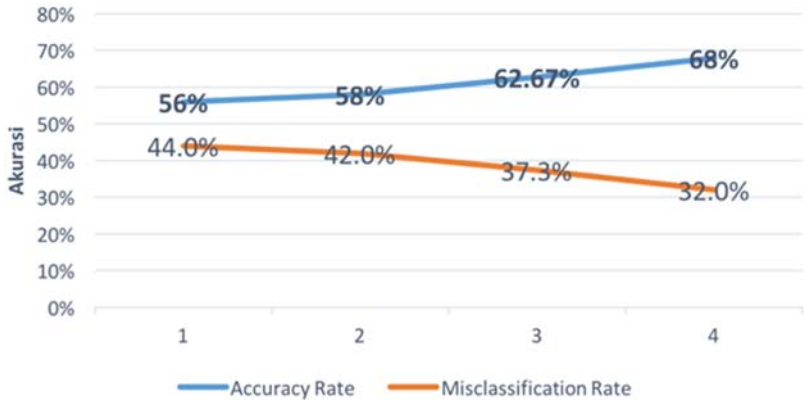
Setelah diketahui data *training* dan data *testing*, kemudian dihitung bobot masing-masing kata menggunakan TF-IDF sebelum melanjutkan ke proses klasifikasi menggunakan algoritma KNN.

TABEL 4
KEMIRIPAN DATA

Data Training	Kemiripan	Klasifikasi
T1	0,554	Travel
T2	0,943	Travel
T3	1,051	Kuliner
T4	1,050	Kuliner
T5	0,851	Olahraga
T6	0,910	Olahraga

TABEL 5
KEMIRIPAN DATA SETELAH DIURUTKAN

Data Training	Kemiripan	Klasifikasi
T3	1,051	Kuliner
T4	1,050	Kuliner
T2	0,943	Travel
T6	0,910	Olahraga
T5	0,851	Olahraga
T1	0,554	Travel



Gambar 6. Grafik Akurasi

Dari proses TF-IDF diketahui bobot masing-masing kata pada semua data *tweet*. Setelah melalui tahap atau proses pembobotan dokumen, maka dapat dilanjutkan ke proses klasifikasi dengan algoritma K-Nearest Neighbor. Algoritma K-Nearest Neighbor diawali dengan tahap menghitung jarak antara bobot setiap kata pada data *testing* dan data *training*, lalu mengurutkan jarak dari jarak yang terdekat hingga yang terjauh. Untuk menentukan hasil klasifikasi didapatkan dari kelas yang paling banyak menjadi tetangga terdekat.

Dari hasil perhitungan *cosine similarity* keenam data, diperoleh hasil yang tertera pada Tabel 4. Data tersebut kemudian diurutkan mulai dari nilai terbesar hingga nilai terkecil seperti ditunjukkan oleh Tabel 5.

Proses yang akan dilakukan setelah pengurutan adalah menentukan yang data yang mana saja yang merupakan tetangga terdekat dari data *test* terhadap data *training*. Jumlah tetangga terdekat adalah sejumlah K, untuk nilai K sudah ditentukan atau diasumsikan sebelumnya sesuai dengan syarat yaitu K harus lebih dari satu, nilai K adalah nilai ganjil karena jika diambil K adalah bilang genap akan ada kemungkinan hasil klasifikasi sulit ditentukan karena masing-masing kelas bernilai sama, nilai K lebih dari jumlah kelas dan nilai K tidak melebihi data *training* [9]. Maka dari itu nilai K=3 karena memenuhi syarat-syarat yang sudah disebutkan sebelumnya, lalu jumlah tetangga yang dipilih adalah sebanyak 3 dokumen.

Pengklasifikasian diambil dari nilai klasifikasi yang sering muncul dan termasuk bagian dari tetangga terdekat. Dapat dilihat pada Tabel 4, kelas yang sering muncul dari tetangga terdekat adalah kelas Kuliner meskipun data kelas Travel muncul namun banyaknya kelas Kuliner lebih banyak muncul dibandingkan kelas Travel, maka kelas yang didapat pada data *testing* atau data baru adalah Kuliner.

TABEL 6
TETANGGA TERDEKAT

Data Training	Kemiripan	Tetangga	Klasifikasi
T3	1,051	Ya	Kuliner
T4	1,050	Ya	Kuliner
T2	0,943	Ya	Travel
T6	0,910	Tidak	Olahraga
T5	0,851	Tidak	Olahraga
T1	0,554	Tidak	Travel

Pengujian Akurasi

Pengujian akurasi yang akan dilakukan adalah dengan menguji hasil keluaran dari algoritma KNN menggunakan metode *confusion matrix* yang tujuannya adalah mencari akurasi dan *error rate* dari algoritma KNN terhadap kasus pada penelitian ini.

Jumlah data uji yang digunakan akurasi dilakukan bertahap mulai dari mulai dari 25, 50 hingga 100 data uji dan data latih yang tersedia adalah 1054 data latih.

Grafik dari empat kali pengujian dengan data uji 25, 50 ,75, 100 dapat dilihat pada Gambar 6. Sebagaimana ditunjukkan oleh Gambar 6 tersebut, nilai akurasi terbesar yang didapat pada proses klasifikasi untuk empat kali pengujian adalah 68%.

4. Kesimpulan

Berdasarkan hasil implementasi dan pengujian pada aplikasi Twikipedia maka diperoleh kesimpulan sebagai berikut:

1) Aplikasi ini dapat membantu memudahkan pengguna Twitter yang menggunakannya sebagai media pemasaran atau promosi untuk melakukan promosi dengan cara melakukan *tweet* promosi terhadap *follower* yang sudah diklasifikasikan.
2) Aplikasi ini dapat membantu memudahkan pengguna Twitter yang menggunakannya setabbagai media pemasaran atau promosi un-

tuk menentukan waktu *tweet* pormosi yang tepat yang sudah dianalisa dari waktu aktif *follower* pengguna.

Referensi

[1] F. Tjiptono dan G. Chandra, Pemasaran Strategik, Yogyakarta: ANDI, 2012.

[2] A. Priyandana, "Marketing Buzz Bikin Interaksi Makin Menggigit," 21 Maret 2015. [Online]. Available: https://gintong.me. [Diakses 6 Mei 2016].

[3] Kementerian Komunikasi dan Informatika Republik Indonesia, "Kominfo: Pengguna Internet di Indonesia 63 Juta Orang," 7 Juli 2013. [Online]. Available: kominfo.go.id. [Diakses 25 April 2016].

[4] We Are Social LTD, "Social media now more popular than TV," 27 Maret 2013. [Online]. Available: wearesocial.com. [Diakses 25 April 2016].

[5] SocialTimes, "Snapchat Is the Fastest Growing Social Network (Infographic)," 28 Juli 2015. [Online]. Available: adweek.com. [Diakses 25 April 2016].

[6] Lembing.com, "Cara Meningkatkan Social Media Engagement Bagi Perusahaan," 12 September 2014. [Online]. Available: http://lembing.com. [Diakses 26 April 2016].

[7] R. Feldmen dan J. Sanger, The Text Mining Handbook, New York: Cambridge University Press, 2007.

[8] R. Feldman dan J. Sanger, The Text Mining Handbook, New York: Cambridge University Press, 2007.

[9] J. Han, M. Kamber dan J. Pei, Data Mining Concepts and Techniques, Waltham: Elsevier, 2012.

[10] C. Freitas, J. Carvalho, J. J. Oliveira, S. Aires dan R. Sabourin, "Confussion Matrix Disagreement for Multiple Classifiers," CIARP, pp. 387-396, 2007.