KLASIFIKASI DOKUMEN KOMENTAR PADA SITUS YOUTUBE MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR* (K-NN)

Moh Aziz Nugroho[1], Heru Agus Santoso[2]
Teknik Informatika, Fakultas Ilmu Komputer,Universitas Dian Nuswantoro
Jl. Nakula I No. 5-11, Jawa Tengah 50131
E-mail : 111201105980@mhs.dinus.ac.id[1],herezadi@gmail.com[2]

Abstrak

Di era serba maju seperti ini persaingan dalam produksi film sangatlah ketat, kemajuan teknologi menjadikan konsumen dapat dengan mudah mendapatkan informasi tentang film yang mereka inginkan misalkan pada situs youtube. Rumah produksi film dituntut untuk dapat memproduksi film dalam waktu singkat dan jumlah banyak agar tidak tertinggal oleh rumah produksi film lainnya karena konsumen menginginkan sebuah film yang selalu baru setiap saat. Kemudian masalah timbul karena rumah produksi film tersebut lebih mementingkan jumlah film yang diproduksi dibandingkan dengan kualitas film tersebut. Sistem klasifikasi komentar diharapkan dapat membantu untuk mengetahui respon positif dan negatif dari pengguna situs youtube yang memberikan komentarnya. Sentiment analysis digunakan untuk mengetahui sikap seseorang dalam konteks dokumen. Sentiment analysis memiliki tahapan preprocessing yang terdiri dari case folding, stopword removal, cleansing, tokenizing, stemming. Pembobotan kata yang digunakan adalah term frequency – invers document frequency dan perhitungan similaritasnya menggunakan cosine similarity kemudian menggunakan k-nearest neighbor sebagai metode klasifikasinya. Hasil yang didapatkan dari implementasi metode k-nn ini cukup baik dengan uji coba sebanyak 6 kali. Rata-rata accuracy tertinggi adalah 0,806 dengan recall 0,848, precission 0,788 dan error rate 0,150. Sedangkan accuracy terendah adalah 0,669 dengan recall 0,88, precission 0,177 dan error rate 0,331.

Kata Kunci: *Sentiment Analysis, K-Nearest Neighbor*

Abstract

In this modern era, competition in movie production is very tight. Technological advance has helped consumers to easily acquire their desired movie information, such as in a site like Youtube. Production houses are required to be able to produce a large quantitiy of movies in quite a short time in order to keep them from being left behind by other production houses. This is due to the need of consumers who want a new movie every time. Problem then arises because the production houses tend to give more concern on movies quantity than movies quality. Comments classification system is expected to help determining both positive and negative responds from Youtube users who provide their comments. Sentiment analysis is utilized to recognize someone's attitude in the document context.Sentiment analysis has a preprocessing stage consisting of case folding, stopword removal, cleansing, tokenizing, stemming. Words weighting used is the term frequency - inverse document frequency and similarity calculation using the cosine similarity then using K-nearest neighbor as a method of classification.Result obtained from the implementation of K-NN method is good enough with as many as six times testing. The average of highest accuracy is 0,806 with as much as 0,848 of recall, 0.788 of precission and 0.150 of error rate. While the lowest accuracy is 0.669 with as much as 0.88 of recall , 0.177 of precission and error rate of 0.331

Keywords: *Sentiment Analysis, K-Nearest Neighbor*

1. PENDAHULUAN

Latar Belakang

Film merupakan salah satu media hiburan bagi masyarakat luas. Film sendiri dapat juga berarti sebuah industri, yang mengutamakan eksistensi dan ketertarikan cerita yang dapat mengajak banyak orang terlibat. Semakin berkembangnya industri perfilman dalam negeri maupun luar negeri dari berbagai rumah produksi, berbanding lurus dengan persaingan dalam menghasilkan karya film yang menarik dan berkualitas untuk para penikmat film diseluruh dunia. Dengan kemajuan teknologi yang pesat sekarang ini, seluruh informasi tentang film–film tersebut sudah tersedia di internet. Semakin banyak informasi yang tersedia di internet, maka akan semakin sulit juga untuk menemukan informasi yang sesuai dengan kebutuhan kita. Jika informasi tersebut diolah dengan baik maka akan di dapatkan nilai tambah dari informasi tersebut, misalkan saja mengetahui apa yang di pikirkan orang lain dari informasi yang ada di internet tersebut [1].

Akan muncul masalah, apabila seseorang atau sebuah organisasi ingin mengetahui respon masyarakat terhadap sebuah film yang tersedia di internet, misalkan di situs Youtube. Youtube adalah situs berbagi video yang memungkinkan pengguna untuk berbagi video yang dimilikinya[2]. Banyak pengguna mengunggah sebuah video atau cuplikan film (*trailer*) yang dimilikinya, kemudian pengguna lain dapat memberikan tanggapan tentang film tersebut dalam bentuk opini berupa pengalaman baik maupun buruk. Dari opini tersebut apabila kita olah dengan baik akan didapatkan sebuah manfaat yang berguna.

Sentiment analysis merupakan bagian dari *opinion mining*, adalah proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi [3]. Dilakukan untuk mengetahui sikap seorang pembicara atau penulis berkaitan dengan topik yang dibahas dari komentar yang di berikan. Komentar dari seseorang yang dipengaruhi oleh emosional penulis (*sentiment analysis*) nantinya akan diklasifikasikan ke dalam kelompok sentiment baik itu positif atau negatif. Salah satu algoritma untuk mengklasifikasikan komentar positif atau negatif adalah algoritma *K-Nearest Neighbor (K-NN). K-NN* adalah suatu metode yang menggunakan algoritma *supervised* dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *K-NN*. Tujuan dari algoritma ini adalah mengklasifikasi objek baru berdasakan atribut dan sampel latih. Diberikan titik uji, akan ditemukan sejumlah K objek (titik *training*) yang paling dekat dengan titik uji. Klasifikasi menggunakan *voting* terbanyak di antara klasifikasi dari K objek. Algoritma *K-NN* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sample uji yang baru [4].

Penggunaan *K-NN* dalam penelitian bidang analisis sentimen telah banyak digunakan, salah satunya ialah untuk mengkategorikan keluhan-keluhan yang diterima oleh salah satu universitas di Indonesia melalui sebuah fasilitas *E-Complaint* yang dimilikinya. Karena begitu banyak data masuk dan SDM yang menanganinya terbatas, maka diperlukan pengelompokan data-data tersebut sehingga tau mana yang diprioritaskan terlebih dahulu [5].

Berdasarkan gambaran di atas, diperlukan adanya sebuah sistem yang dapat mengklasifikasikan review film berbahasa Inggris ke dalam dua *sentiment* yaitu positif dan negatif pada review film atau *trailer bergenre* aksi

di situs youtube.com guna mempermudah user mengetahui kualitas dari film yang akan dinikmati dikarenakan terdapat banyak film yang diproduksi hanya untuk bersaing dengan rumah produksi film lainnya tanpa memperdulikan kualitas film itu sendiri. Kemudian pada penelitian ini penulis menggunakan *Key-Nearest Neighbor (K-NN)* sebagai metode klasifikasi tersebut.

Rumusan Masalah

Berdasarkan pada permasalahan yang telah dijelaskan pada bagian latar belakang, maka rumusan masalah dapat disusun sebagai berikut:

1. Bagaimana implementasi dari pengembangan *sentiment analysis* pada review film *bergenre* aksi di youtube dengan metode *K-Nearest Neighbor*.
2. Berapa tingkat akurasi yang dihasilkan dari percobaan pada penerapan klasifikasi menggunakan metode *K-Nearest Neighbor* ini.

Tujuan

Berdasarkan pada permasalahan yang telah dijelaskan pada bagian latar belakang, maka rumusan masalah dapat disusun sebagai berikut:

1. Bagaimana implementasi dari pengembangan *sentiment analysis* pada review film *bergenre* aksi di youtube dengan metode *K-Nearest Neighbor*.
2. Berapa tingkat akurasi yang dihasilkan dari percobaan pada penerapan klasifikasi menggunakan metode *K-Nearest Neighbor* ini.

Manfaat

Dari hasil penelitian ini diharapkan akan memberikan beberapa manfaat kepada pembaca dan penulis. Manfaat yang diharapkan adalah sebagai berikut:

1. Bagi Penulis
   a. Menerapkan ilmu yang diperoleh dari Program Teknologi Informatika Universitas Dian Nuswantoro Semarang
   b. Mendapatkan pengetahuan lebih dalam tentang perancangan dan pengembangan *sentiment analysis* pada review film dengan metode *K-NN*.
2. Bagi Pengguna
   a. Memudahkan pengguna mengetahui lebih awal kualitas film yang akan dilihat.
   b. Dengan mengetahui jenis film yang mendapat banyak respon positif akan memudahkan produser dalam memproduksi film.

Text Mining

Text mining merupakan bagian dari penerapan data mining yang digunakan untuk mencari pola dalam teks. Berbeda dengan data mining yang menggunakan data biasa, dalam *text mining* ini yang digunakan adalah data dalam bentuk dokumen atau teks. *Text mining* ini bertujuan untuk menemukan pola dalam sebuah teks atau dokumen yang sebelumnya tidak terlihat menjadi pola yang diinginkan untuk tujuan tertentu [1].

Sentimen Analisis

Analisis sentimen merupakan bagian dari opinion mining, adalah proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi[3]. Dilakukan untuk mengetahui sikap seorang pembicara atau penulis berkaitan dengan topik yang dibahas dari komentar yang di berikan. Komentar dari seseorang yang

dipengaruhi oleh emosional penulis (*sentiment analysis*) nantinya akan dikelompokkan ke dalam kelompok sentiment baik itu positif maupun negatif.

Text Preprocessing

Dikarenakan dokumen teks memiliki data yang tidak terstruktur maka digunakanlah *text processing* ini untuk merubah data yang belum terstruktur itu menjadi sebuah data yang terstruktur sehingga dapat siap untuk digunakan dalam proses selanjutnya. *Text Preprocessing* ini memiliki beberapa tahapan yaitu:

a. Mengekstrak teks yang akan kita olah.
b. Melakukan *stopword removal*, yaitu menghilangkan kata-kata yang tidak bermakna misalkan kata hubung.
c. Melakukan *stemming*, yaitu menghilangkan imbuhan-imbuhan dalam sebuah kata, singkatnya ialah merubah kata berimbuhan menjadi kata dasar.

Term Frequency-Inverse Document Frequency (TF-IDF)

TF *(Term Frequency)* ialah frekuensi kemunculuan kata pada setiap dokumen, dari TF tersebut didapatkan DF *(document frequency)* yaitu banyaknya dokumen yang mengandung suatu kata tersebut.TF-IDF merupakan sebuah nilai yang digunakan untuk menghitung bobot sebuah kata yang muncul dalam dokumen. TF-IDF didapatkan dari hasil perkalian antar TF dan IDF, dimana IDF merupakan hasil invers dari DF. Perhitungannya dapat dituliskan sebagai berikut [7]:

$$\mathrm{IDF(w)} = \log\left(\frac{\mathrm{N}}{\mathrm{DF(w)}}\right) \quad (1)$$

$$\mathrm{TF-IDF(w,d)} = \frac{\mathrm{TF-IDF(w,d)}}{\sqrt{\sum_{w=1}^{n} \mathrm{TF-IDF(w,d)}^2}} \quad (2)$$

Keterangan:
IDF(w) : bobot kata dalam seluruh dokumen
w : sebuah kata
TF(w,d) : frekuensi kemunculan kata *w* dalam dokumen *d*
IDF(w) : inverse *DF* dari kata *w*
N : jumlah seluruh dokumen
DF(w) : jumlah dokumen yang mengandung kata *w*

Kemudian untuk menstandarisasi nilai TF-IDF ke dalam interval 0 sampai 1 diperlukan normalisasi, rumus (3) dinormalisasi dengan rumus (4) sebagai berikut [7]:

$$\mathrm{TF-IDF(w,d)} = \frac{\mathrm{TF-IDF(w,d)}}{\sqrt{\sum_{w=1}^{n} \mathrm{TF-IDF(w,d)}^2}} \quad (3)$$

Dimana $\sum_{w=1}^{n} \mathrm{TF-IDF(w,d)}$ adalah jumlah dari nilai TF-IDF dari kata pertama hingga kata ke n yang terdapat dalam dokumen d.

Cosine Similarity

Cosine similarity berfungsi untuk membandingkan kemiripan antar dokumen, dalam hal ini yang dibandingkan adalah *query* dengan dokumen latih [3]. Rumus dapat dilihat sebagai berikut [7]:

$$\mathrm{cosSim}(d_j, q_k) = \frac{\sum_{i=1}^{n}(td_{ij} \times tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij}^2 \times \sum_{i=1}^{n} tq_{ik}^2}} \quad (4)$$

Keterangan:
cosSim(dj,qk) : tingkat kesamaan dokumen dengan query tertentu
tdij : term ke-i dalam vektor untuk dokumen ke-j
tqik : term ke-i dalam vektor untuk *query* ke-k
n : jumlah *term* yang unik dalam data set

K-Nearest Neighbor (K-NN)

Algoritma *k-nearest neighbor* (K-NN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *K-NN* termasuk algoritma *supervised learning* dimana hasil dari *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada *K-NN*. Kelas yang paling banyak muncullah yang akan menjadi hasil klasifikasi. Klasifikasi menggunakan *voting* terbanyak diantara klasifikasi dari k obyek.. algoritma *k-nearest neighbor (K-NN)* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari *query instance* yang baru [4].

Perhitungannya adalah dengan menjumlahkan semua nilai kemiripan yang tergolong dalam satu kategori kemudian membandingkan manakah yang lebih besar. Rumusnya adalah sebagai berikut:

$$\mathrm{p}(x,\mathrm{c_m}) = \sum_{l=1}^{m} SIM\left(X, d_j\right) \in \mathrm{c_m} \quad (5)$$

Keterangan:

P(x,cm) : probabilitas dokumen X menjadi anggota kategori cm

sim(x,dj)∈cm : kemiripan antara dokumen X dengan dokumen latih dj yang merupakan anggota dari kategori cm

m : jumlah sim(x,dj) yang termasuk dalam kategori cm

Confusion Matrix

Confusion matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi pada bidang data mining. *Confusion matrix* ini nantinya akan melakukan perhitungan yang melakukan 4 keluaran, yaitu *recall* (proporsi kasus positif yang diidentifikasi dengan benar), *precision* (proporsi kasus dengan hasil positif yang benar, *accuracy* (perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus) dan *eror rate* (kasus yang diidentifikasi salah dengan jumlah seluruh kasus [9]. Rumus metode ini adalah sebagai berikut:

$$\text{recall} = \frac{d}{c+d} \quad (6)$$

$$\text{precision} = \frac{d}{(b+d)} \quad (7)$$

$$\text{accuracy} = \frac{(a+d)}{(a+b+c+d)} \quad (8)$$

$$\text{error rate} = \frac{(b+c)}{a+b+c+d} \quad (9)$$

Keterangan:

a = jika hasil klasifikasi (-) dan data asli (-)

b = jika hasil klasifikasi (+) dan data asli (-)

c = jika hasil klasifikasi (-) dan data asli (+)

d = jika hasil klasifikasi (+) dan data asli (+)

2. METODE

Kebutuhan Software

Adapun perangkat lunak yang dibutuhkan dalam penelitian ini adalah sebagai berikut :

a. Sistem Operasi
   Sistem operasi yang digunakan dalam penelitian ini adalah Windows 7.
b. *Hypertext Preprocessor (PHP)*
   Merupakan bahasa pemrograman yang akan digunakan untuk mengimplementasi hasil dari rancangan yang sudah dibuat.

c. Notepad++
   Tools yang digunakan untuk editor bahasa pemrograman PHP.
d. MySql
   Software ini digunakan untuk menyimpan database dari sistem yang akan dibangun.
e. Ms. Word
   Software ini digunakan untuk membuat laporan hasil penelitian.

Kebutuhan *Hardware*

Selain kebutuhan *software*, diperlukan pula *hardware* yang harus dipenuhi agar penelitian ini berjalan dengan lancar. Adapun *hardware* yang digunakan dalam penelitian ini adalah sebagai berikut:

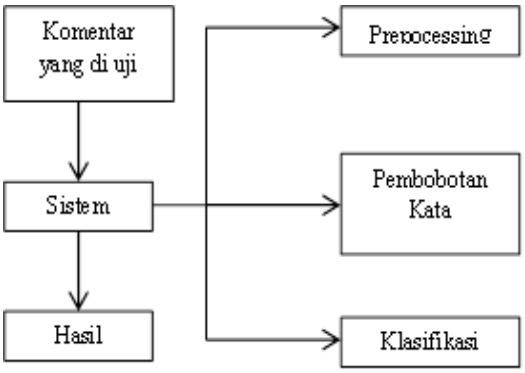
a. *Personal Computer* atau laptop dengan spesifikasi :
   Prosesor : Dual core
   Sistem Operasi : Windows 7
   RAM : 2 GB
b. Printer, digunakan untuk mencetak hasil penelitian ke dalam bentuk *hardcopy*.

Pengumpulan Data

Untuk mendapatkan data yang nantinya akan digunakan dalam penelitian ini, penulis mendapatkannya dari peneliti terdahulu dari *cornell university* yang sudah mengelompokkan beberapa komentar ke dalam kelompok positif dan negatf. Selain itu juga penulis mengumpulkan sendiri dari komentar-komentar yang diberikan oleh pengguna situs youtube.

Metode yang diusulkan

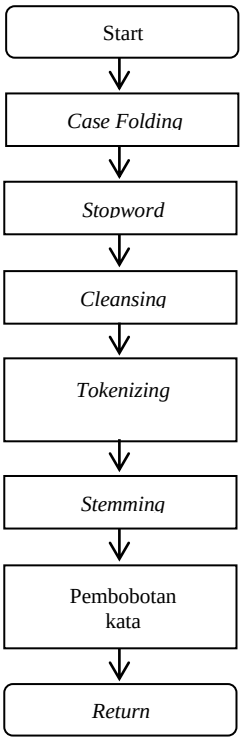
Proses *text mining* secara umum memiliki tahapan yaitu *preprocessing text,* kemudian dilakukan pembobotan kata dan diolah menggunakan algoritma yang dipakai dalam kasus ini menggunakan algoritma *K-NN*. *Preprocessing text* juga terdiri dari beberapa tahapan yaitu *cleansing, parsing, tokenizing, stopword removal, stemming*, dan pembobotan kata. Alur prosesnya jika digambarkan adalah sebagai berikut:



Gambar 1. Rancangan Arsitektur Sistem
[Sumber: Analisa Penulis]

Preprocessing

Karena sistem tidak bisa membaca dokumen teks dikarenakan strukturnya tidak teratur maka diperlukanlah tahapan prepocessing yaitu merubah dari teks menjadi sebuah angka yang terstruktur sehingga dapat dikenali oleh sistem. Prepocessing ini terdiri dari beberapa langkah yaitu *cleansing, parsing, tokenizing, stopword removal, stemming* dan pembobotan kata. Adapun diagram alirnya adalah sebagai berikut:



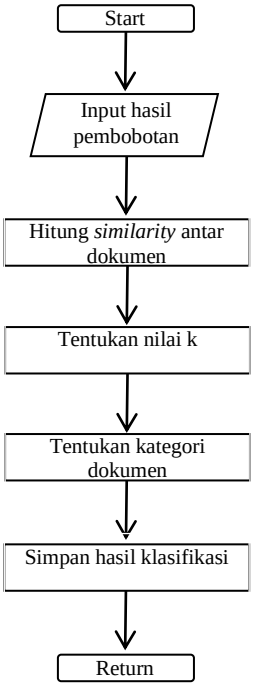
Gambar 2. Diagram Alir Prepocessing
[Sumber: Analisa Penulis]

Langkah pertama yang dilakukan adalah *case folding*, yaitu merubah kalimat tersebut ke dalam bentuk huruf kecil seluruhnya, mengganti huruf kapital menjadi huruf kecil agar seragam. Kemudian setelah kalimat tersebut seragam maka selanjutnya adalah melakukan penyaringan kata-kata yang tidak bermakna. Kata-kata yang tidak bermakna ini akan memakan banyak memori dan pemborosan waktu proses. Kata-kata yang disaring nantinya yaitu kata hubung misalnya tahap ini dinamakan *stopword removal*. Pada tahap *cleansing* akan dilakukan pembersihan dokumen dari tanda baca dan simbol-simbol. Tanda baca dan simbol-simbol itu akan dideteksi dan digantikan oleh spasi. Kemudian dilakukan *tokenizing*, dokumen yang sudah dibersihkan dari tanda baca dan simbol-simbol nantinya akan dipecah menjadi kata, pemecahan ini berdasarkan oleh spasi. Langkah selanjutnya yaitu melakukan *stemming*, yaitu merubah kata-kata tersebut menjadi kata dasar. Proses ini akan menghilangkan imbuhan kata, yakni awalan, sisipan, awalan-akhiran.

Karena komputer tidak dapat membaca dokumen teks, maka perlulah representasi dari teks menjadi angka, tahap ini yaitu pembobotan kata.Tiap kata yang muncul dalam dokumen tadi akan di beri bobot tergantung dari frekuensi kemunculannya dalam tiap dokumen. Bobot ini akan disimpan dan digunakan untuk langkah perhitungan menggunakan algoritma *K-NN*.

Klasifikasi Menggunakan *K-Nearest Neighbor (K-NN)*

Pada tahap klasifikasi ini yaitu mengambil hasil dari pembobotan kata. Hasil dari pembobotan kata tersebut selanjutnya akan dihitung nilai similaritasnya atau kemiripan antara dokumen uji dengan dokumen latih menggunakan rumus (4). Setelah didapatkan nilai similaritasnya tentukan nilai k, ambil hasil similaritas tersebut sesuai nilai k di mulai dari nilai similaritas yang paling tinggi. Setelah mendapatkan hasil similaritas sejumlah nilai k, tentukan hasil klasifikasi menggunakan rumus (5. Berikut adalah diagram alirnya:



Gambar 3. Diagram Alir Proses K-NN
[Sumber: Analisa Penulis]

3. HASIL DAN PEMBAHASAN

Untuk mengetahui nilai akurasi dari penerapan metode k-nn ini, penulis mencoba melakukan beberapa uji coba dengan enam skenario, yaitu dengan menggunakan data uji yang sama pada tiap skenario tetapi dengan data *training* yang berjumlah berbeda antara skenario satu dengan yang lainnya. Kemudian menghitung nilai akurasi pada tiap skenario dengan *confusion matrix.*

Analisa *confusion matrix* disini penulis menitikberatkan pada nilai *accuracy* dan *eror rate*, semakin besar nilai *accuracy* maka semakin akurat klasifikasi dari penerapan metode k-nn ini, sebaliknya semakin tinggi nilai *eror rate* maka semakin rendah akurasi dari penerapan metode k-nn ini.

Skenario 1

Uji coba skenario pertama penulis menggunakan data uji sebanyak 32 dan data *training* 120 dengan porsi 60 berlabel positif dan 60 berlabel negatif.

Tabel 1. Uji Coba Skenario 1

Nilai K	Confusion matrix			
	recall	precision	accuracy	Error rate
2	0,647	0,647	0,625	0,375
10	0,647	0,765	0,75	0,25
15	0,867	0,867	0,813	0,187
35	0,813	0,765	0,781	0,219
50	0,867	0,765	0,844	0,156
Rata-rata	0,768	0,762	0,763	0,237

[Sumber: Hasil Analisa]

Dari skenario 1 yang mendapatkan hasil pada table 4.1 dapat diketahui nilai terendah dari *accuracy* adalah 0,625 pada k=2 dan nilai tertinggi 0,844 pada k=50. Sedangkan nilai *eror rate* terendah adalah 0,156 pada k= 50 dan nilai tertinggi 0,375 pada k=2.

Skenario 2

Pada uji coba skenario 2 menggunakan data uji yang sama dan data *training* sebanyak 180 dengan porsi 90 berlabel positif dan 90 berlabel negatif. Pada skenario 2 ini mendapatkan hasil berikut:

Table 2. Uji Coba Skenario 2

Nilai K	Confusion matrix			
	recall	precision	accuracy	Error rate
2	0,643	0,529	0,594	0,406
10	0,765	0,765	0,75	0,25
15	0,786	0,647	0,719	0,219
35	0,929	0,765	0,843	0,156
50	1	0,722	0,844	0,156
Rata-rata	0,825	0,686	0,75	0,237

[Sumber: Hasil Analisa]

Hasil dari skenario 2 yang terdapat tabel pada 2 mempunyai hasil mirip dengan skenario 1, nilai tertinggi *accuracy* terdapat pada k= 50 yaitu 0,844 dan nilai terendah pada k= 2 adalah 0,594. Begitu pula pada nilai *error rate*, nilai

terendah pada k= 50 dan k= 35 yaitu 0,156 dan nilai tertinggi pada k= 2 adalah 0,406.

Skenario 3

Skenario 3 penulis masih menggunakan data uji yang sama dan data *training* sebanyak 210 dengan porsi 90 berlabel positif dan 120 berlabel negatif. Skenario 3 ini mendapatkan hasil berikut:

Tabel 3. Uji Coba Skenario 3

Nilai K	Confusion matrix			
	recall	precision	accuracy	Error rate
2	0,75	0,667	0,688	0,313
10	0,652	0,882	0,688	0,313
15	0,7	0,875	0,75	0,25
35	0,8	0,941	0,844	0,156
50	0,727	0,941	0,781	0,219
Rata-rata	0,726	0,861	0,750	0,25

[Sumber: Hasil Analisa]

Pada skenario ke 3 ini mendapatkan hasil yang lebih baik dari sebelumnya yaitu nilai tertinggi 81,579% dan nilai terendah 55,263% yang mempunyai rata-rata sebanyak 70,526%.

Skenario 4

Kemudian pada skenario 4 ini penulis mencoba menggunakan data *training* dengan jumlah yang sama dengan skenario 3 tetapi dengan porsi yang berbeda, yaitu 120 dokumen berlabel positif dan 90 berlabel negatif.

Tabel 4. Uji Coba Skenario 4

Nilai K	Confusion matrix			
	recall	precision	accuracy	Error rate
2	0,692	0,529	0,625	0,375
10	0,818	0,5	0,656	0,344
15	0,889	0,444	0,656	0,344
35	1	0,471	0,719	0,281
50	1	0,412	0,688	0,313
Rata-rata	0,88	0,177	0,669	0,331

[Sumber: Hasil Analisa]

Walaupun jumlah dokumen *training* sama dengan skenario 3, ternyata dalam skenario 4 mendapatkan hasil yang berbeda. Pada skenario 4 ini nilai tertinggi *accuracy* mencapai angka 0,719 dan nilai terendah 0,625. Sedangkan nilai *error rate* tertinggi yaitu 0,375 dan nilai terendah 0,281. Skenario 3 dan 4 ini membuktikan pengaruh dari penentuan jumlah porsi yang tepat dari setiap label.

Skenario 5

Pada skenario 5 ini, penulis menambahkan jumlah dokumen *training* yaitu 240 dokumen dengan porsi seimbang yaitu 120 berlabel positif dan 120 berlabel negatif. Uji coba skenario 5 ini mendapatkan hasil seperti pada tabel 5 dibawah ini:

Tabel 5. Uji Coba Skenario 5

Nilai K	Confusion matrix			
	recall	precision	accuracy	Error rate
2	0,706	0,706	0,688	0,313
10	0,789	0,882	0,813	0,188
15	0,813	0,765	0,781	0,219
35	0,933	0,824	0,875	0,125
50	1	0,765	0,875	0,125
Rata-rata	0,848	0,788	0,806	0,150

[Sumber: Hasil Analisa]

Dapat dilihat dengan menambahkan jumlah dokumen *training* ternyata nilai *accuracy* dan *error rate* meningkat dengar rata-rata yang baik. *Accuracy* tertinggi yaitu 0,875 pada k=35 dan 50, sedangkan nilai terendahnya adalah 0,688 pada k=2 dengan rata-rata 0,806. Kemudian error rate tertinggi adalah 0,313 pada k= 2, sedangkan nilai terendahnya adalah 0,125 pada k= 50 dengan rata-rata 0,150.

Skenario 6

Skenario terakhir penulis mencoba menggunakan data yang sama dengan skenario 5 yaitu 120 dokumen berlabel

positif dan 120 berlabel negatif, yang membedakan adalah pada skenario 6 ini penulis mencoba menggunakan pengujian tanpa *stemming*.

Tabel 6. Uji Coba Skenario 6

Nilai K	Confusion matrix			
	recall	precision	accuracy	Error rate
2	0,667	0,706	0,656	0,343
10	0,813	0,765	0,781	0,219
15	0,869	0,765	0,813	0,188
35	0,929	0,765	0,844	0,156
50	1	0,765	0,875	0,125
Rata-rata	0,856	0,753	0,794	0,206

[Sumber: Hasil Analisa]

Dari tabel 4.6 diketahui ternyata *stemming* mempunyai peranan dalam akurasi penelitian ini walaupun tidak terlalu signifikan. Dari skenario 6 didapatkan hasil accuracy tertinggi adalah 0,875 dan nilai terendahnya 0,656. Sedangkan error rate tertinggi adalah 0,343 dan nilai terendahnya 0,125.

4. KESIMPULAN DAN SARAN

Kesimpulan

Kesimpulan yang didapatkan dari hasil skripsi klasifikasi dokumen komentar film di youtube menggunakan metede k-nearest neighbor ini adalah sebagai berikut:

1. Jumlah dokumen *training* dan nilai k sangat berpengaruh dalam pengklasifikasian. Dengan menggunakan dokumen *training* dan nilai k yang tepat, maka akan didapatkan hasil klasifikasi yang baik.
2. Hasil pengujian klasifikasi dokumen komentar pada situs youtube menggunakan metode k-nn ini memiliki nilai accuracy tertinggi yaitu dengan rata 0,806 dengan jumlah dokumen *training* 240 menggunakan proses *stemming*. Sedangkan *accuracy* terendahnya adalah 0,669 dengan dokumen *training* 210 yang mempunyai porsi 120 dokumen berlabel positif dan 90 berlabel negatif. Untuk pengujian tanpa stemming memiliki rata-rata accuracy hingga 0,794.
3. Klasifikasi dengan metode *k-nearest neighbor* ini dapat diterapkan dengan otomatis yang mempunyai langkah yaitu preprocessing terlebih dahulu untuk dapat melakukan pembobotan kata dan menghitung similaritas atau *cosine similarity*, setelah itu diambil nilai cosine sebanyak nilai k yang dimasukkan user untuk mengklasifikasikan dokumen tersebut.

Saran

Saran yang bisa penulis berikan dari hasil skripsi klasifikasi dokumen komentar pada youtube dengan metode k-nn ini adalah:

1. Mencoba menggunakan algoritma stemming yang lain, dikarenakan *stemming porter* masih belum bisa membuat kata dasar yang akurat, misal *stupid* dan *stupidest* masih ada dalam satu proses.
2. Memperhatikan persamaan kata yang muncul agar bisa dilakukan pembersihan karena pengklasifikasian ini menghitung kemiripan kata yang muncul antar dokumen.
3. Bisa dilakukan pengembangan dari klasifikasi dokumen komentar ini menggunakan metode lainnya atau bisa juga melakukan penggabungan dengan metode k-nn ini.

DAFTAR PUSTAKA

[1] Feizar, Faldy Hildan. 2014. "Analisis Sentimen Opini Film Berbahasa Indonesia Berbasis Kamus Menggunakan Metode *Neighbor-*

Weighted K-Nearest Neighbor", Teknik Informatika. Universitas Brawijaya

[2] Youtube. 2014. "Tentang Youtube", http://www.youtube.com/yt/about/id/index.html [01 Januari 2015]

[3] Luhulimu, Yugo Yudansha. 2013. "*Sentiment Analysis* pada *Review* Barang Bahasa Indonesia dengan Metode *K-Nearest Neighbor* (K-NN)". Teknik Informatika. Universitas Brawijaya.

[4] Permana, Sigit Budi. 2012. "Pengertian, Kelebihan dan Kekurangan *K-Nearest Neighbor* (K-NN)". http://cgeduntuksemua.blogspot.com/2012/03/pengertian-kelebihan-dan-kekurangan-k.html [05 Januari 2015]

[5] Baharsyah, Imam. 2014. "Klasifikasi *Deep Sentiment Analysis E-Complaint* Universitas Brawijaya Menggunakan Metode *K-Nearest Neighbor".* Teknik Informatika. Universitas Brawijaya

[6] Rozi, Imam Fahrur. 2012. Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tingi. Jurnal EECCIS Universitas Brawijaya Vol. 6, No. 1. Hlm 37-43

[7] Indranandita, Amalia. 2008. "Sistem Klasifikasi dan Pencarian Jurnal dengan Menggunakan Metode Naive Bayes dan Vector Space Model". Jurnal Informatika Universitas Kristen Duta Wacana, volume 4, nomor 2. Hlm 9-17

[8]http://itee.uq.edu.au/%7Einfs4203/Lecture/Lesson07_Text_Mining_2011.pdf

[9] Julianto, Windy.2014. "Menghitung Akurasi dengan Confusion Matrix".http://www.tahuituenak.blogspot.ae/2014/04/menghitung-akurasi-dengan-confusion.html[19 Februari 2016].