Klasifikasi Topik Berita Berbahasa Indonesia Menggunakan *k-Nearest Neighbor*

Andi Ahmad Irfa[1], Adiwijaya[2], Mohamad Syahrul Mubarok3

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]andirfa@student.telkomuniversity.com, [2]adiwijaya@telkomuniversity.ac.id,
[3]msyahrulmubarok@gmail.com

Abstrak

Masyarakat Indonesia kini mulai beralih dari konsumsi berita dalam bentuk surat kabar ke berita *online.* Persentase konsumsi berita melalui *online* mencapai 96 persen berdasarkan riset lembaga Global GFK. Angka tertinggi dibandingkan dengan konsumsi berita melalui televisi sebesar 91 persen, surat kabar 31 persen dan radio sebesar 15 persen. Akan tetapi begitu banyak berita bisa menyulitkan kerja editor dalam mengategorikan setiap berita yang ada, oleh karena itu dibutuhkan suatu sistem yang bisa mengategorikan berita sesuai dengan kategori masing-masing. Pada penelitian ini bertujuan untuk membuat suatu sistem yang mampu mengategorikan setiap berita berdasar dari topik berita tersebut. Metode yang digunakan dalam mengklasifikasi berita adalah *k-Nearest Neighbor (K-NN)* yang merupakan algoritma klasifikasi sederhana namu memiliki performa yang tinggi. Pada penelitian ini perancangan sistem dilakukan proses pengumpulan dataset, *preprocessing data,* klafikasifikasi dengan *k-nn,* dan terakhir dilakukan pengujian system. Dalam penelitian ini system yang dibangun mampu menghasilkan performa *micro average f1-measure* sebesar 69,9% dengan nilai *k=16*.

Kata Kunci: Klasifikasi Teks, *Text Mining, K-Nearest Neighbors*

Abstract

Indonesian society is now starting to roll out from news consumption in the form of newspapers to online news. The percentage of online news consumption reached 96 percent based on Global GFK research institute. The percent number is the reverse of the number of news stories. It will be very much news to complicate the work of editors in categorizing every news that there is, therefore required a system that can categorize the news according to their respective categories. In this study is for a system that is able to categorize any news based on the news topic. The method used in classifying news is k-Nearest Neighbor (K-NN) which is a simple content algorithm namu has high performance. In this research the system design is done the process of completion of dataset, preprocessing data, klafikasifikasi with k-nn, and last done by testing system. In this research the built system is able to produce the average micro performance of f1-measure equal to 69,9% with value k = 16.

Keywords: Classification Text, Text Mining, K-Nearest Neighbors

1. Pendahuluan

1.1 Latar Belakang

Berita merupakan cerita atau keterangan mengenai kejadian atau peristiwa yang hangat [1]. Di era digital sekarang, media *online* telah menjadi sasaran bagi masyarakat pada umumnya untuk mencari berita dan informasi. Masyarakat Indonesia kini mulai beralih dari konsumsi berita melalui surat kabar ke berita *online*. Berdasarkan hasil riset yang dilakukan oleh lembaga global GFK dan Indonesia Digital Association (IDA) yang dilaksanakan di lima kota besar di Indonesia sepanjang tahun 2015, persentase konsumsi berita melalui *online*mencapai 96 persen. Angka tertinggi dibandingkan dengan konsumsi berita melalui televisi sebesar 91 persen, surat kabar 31 persen dan radio sebesar 15 persen dari 1.521 responden [2]. Hal ini dapat dimanfaatkan *media online* untuk terus memberikan berita terkini kepada masyarakat. Akan tetapi begitu banyak berita bisa menyulitkan kerja editor dalam mengategorikan setiap berita yang ada. Setiap editor harus membaca setiap berita, ketelitian dan waktu sangat diperlukan untuk mengategorikan setiap berita. Oleh karena itu dibutuhkan sebuah sistem yang bisa mengategorikan berita sesuai dengan isi berita.

Text mining adalah salah satu dari cabang ilmu penambangan data yang dimana proses penambangan data yang berupa teks dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari tiap dokumen sehingga dapat dilakukan analisa keterhubungan setiap dokumen [3]. Tahapan

yang terdapat pada penambangan data teks adalah *preprocessing data* dan *feature extraction*. Pada tugas akhir ini diterapkan metode klasifikasi data yang tertuju pada pemilihan kategori berita dari sekumpulan berita ke dalam kategori yang telah didefinisikan sebelumnya. Oleh karena itu, pada Tugas Akhir ini menggunakan metode *k-Nearest Neighbor* (*k*-NN) yang merupakan algoritma klasifikasi sederhana namun memiliki performa yang tinggi. [4].

1.2 Topik dan Batasannya

Berdasarkan latar belakang di atas, maka perumusan masalah dalam pembuatan tugas akhir ini adalah bagaimana cara mengklasifikasikan topik berita sesuai dengan kategori menggunakan metode K-NN dan bagaimana menganalisis tingkat performa dari sistem yang dibangun.

Dalam pengerjaan tugas akhir ini terdapat beberapa batasan masalah yaitu data artikel berita yang digunakan adalah data artikel berita yang berbahasa Indonesia. Kemudian kategori berita yang digunakan dalam pembuatan tugas akhir ini sejumlah 12 antara lain wisata, teknologi, olahraga, politik, pendidikan, otomotif, *lifestyle*, kesehatan, kriminal, *entertaiment*, ekonomi dan budaya. Bagian dari artikel berita yang diambil hanya judul dan isi artikel berita yang disimpan dalam bentuk (.txt). Dataset yang dikumpulkan oleh penulis sebanyak 360 data teks dimana setiap kelas memiliki 30 data. Berita yang diambil adalah artikel berita dari rentang tanggal 1 November 2016 – 30 April 2017 dari halaman *website* kompas.com, tribunnews.com, republika.com, mediaindonesia.com dan sindonews.com. Lalu pada sistem klasifikasi yang dibangun untuk menangani masalah *multiclass*, yang diimplementasi dilakukan menggunakan bahasa pemrograman *Python* dengan Pengujian k dari k = 1 sampai dengan k = 30.

1.3 Tujuan

Berdasar dari perumusan masalah di atas, maka tujuan pembuatan Tugas Akhir ini antara lain mengklasifikan berita sesuai dengan kategori masing-masing yang telah ditentukan dengan menggunakan metode klasifikasi *k-nearest neighbors* dan menganalisa tingkat performa sistem yang dibangun dengan metode klasifikasi *k-nearest neighbors.*

1.4 Organisasi Tulisan

1. Pendahuluan

   Pada bab ini dijabarkan mengenai latar belakang, perumusan masalah, tujuan, batasan masalah, metodologi penyeles [5]ain masalah dan sistematika penulisan.

2. Studi Terkait

   Pada bab ini terdapat ringkasan hasil kajian pustaka yang terkait dengan masalah yang diajukan penulis pada bab pendahuluan dan menjelaskan teori-teori pendukung yang digunakan untuk menyelesaikan masalah yang diajukan oleh penulis.

3. Sistem yang Dibangun

   Pada bab ini berisi rancangan sistem yang dibangun, pengumpulan dan ilustrasi pengolahan data yang dilakukan dalam membangun sistem pengklasifikasian artikel berita bahasa Indonesia serta skenario pengujian yang dilakukan.

4. Evaluasi

   Pada bab ini membahas hasil pengujian berdasarkan metode dan skenario pengujian yang telah dituliskan pada bab Perancangan Sistem serta pada bab ini juga menjelaskan hasil analisis terhadap hasil pengujian yang telah dilakukan.

5. Kesimpulan

   Pada bab ini menjelaskan kesimpulan dari keseluruhan hasil pengerjaan tugas akhir yang mengacu pada tujuan penelitian, skenario pengujian dan analisis hasil pengujian pada bab-bab sebelumnya serta memberikan saran untuk penelitian selanjutnya.

2. Studi Terkait

Penelitian pada Tugas Akhir ini termasuk dalam bagian dari klasifikasi teks. Penelitian terkait dengan masalah yang diangkat oleh penulis sudah banyak dilakukan sebelumnya, diantaranya klasifikasi topik lirik lagu dengan metode *multinomial naïve bayes* [6], *Aspect-based sentiment analysis to review producrs using naïve bayes* [7], kategorisasi topik tweet di kota jakarta, bandung, dan makassar dengan metode multinomial naïve bayes classifier [8], analisis sentimen pada ulasan buku berbahasa inggris menggunakan information gain dan support vector machine [9], klasifikasi sentimen pada level aspek terhadap ulasan produk berbahasa inggris menggunakan bayesian network (case study: data ulasan produk amazon) [5], klasifikasi sentimen pada ulasan buku berbahasa inggris

menggunakan information gain dan naïve bayes [10], a comparative study of mfcc-knn and lpc-knn for hijaiyyah letters pronounciation classification system [11]. Pada penelitian tersebut mengklasifikasinya mayoritas menggunakan *naïve bayes naïve bayes* yang dikhusukan untuk *text mining*. Pada penelitian klasifikasi topik lirik lagu dengan metode *multinomial naïve bayes* didapatkan bahwa performa system tersebut diperoleh nilai *f1-measure* sebesar 88.91%.

Pada penelitian Tugas Akhir ini, metode yang digunakan adalah *k-nearest neighbor (k-nn).* Penelitian lain yang menggunakan k-NN adalah Implementasi teks mining dengan k-Nearest Neighbor pada analisis [12]. Proses pada analisis sentimen diawali dengan preprocessing, dilanjutkan dengan ekstraksi cirri Term Frequency – Inverse Document Frequency (TF-IDF), kemudian pengkategorian yang terdiri dari perhitungan cosine similarity dan klasifikasi sentimen. Dalam penelitian ini mampu diperoleh nilai *fi-measure* sebesar 77%. Kemudian pada penelitian analisis sentimen level aspek pada ulasan produk menggunakan *k-Nearest Neighbor* oleh Novi S, Mohamad Syahrul Mubarok, dan Adiwijaya [13]. Dalam penelitian tersebut menggunakan metode ekstraksi fitur dan sentimen *Term Frequency-Invers Document Frequency* (TF-IDF) dan menggunakan metode klasifikasi *K-Nearest Neighbor* (K-NN). Kemudian karena K-NN memiliki waktu komputasi yang cukup lama, maka akan digunakan *Principal Component Analysis* (PCA) untuk mereduksi dimensi. Dan Pada penelitian tersebut menunjukkan bahwa nilai rata-rata akurasi tertinggi sebesar 79.58% dengan parameter uji jumlah *principal components* sebanyak 90 dan nilai k = 31.

2.1. Teks Mining

Text mining adalah salah satu dari cabang ilmu penambangan data yang dimana proses penambangan data yang berupa teks dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari tiap dokumen sehingga dapat dilakukan analisa keterhubungan setiap dokumen [3]. Pada awalnya pembentukan *text mining* ialah untuk mendapatkan bentuk-bentuk informasi yang dapat dianalisis lebih dalam dari suatu teks atau dokumen yang bentuknya belum terstruktur. Di era digital, *text mining* telah menjadi animo dalam berbagai aspek bidang, antara lain dibidang akademik, keamanan, pengembangan perangkat lunak dan aplikasi, *marketing*, media *online*, dan biomedis.

2.2. *Pre-Processing*

Pre-processing merupakan langkah-langkah dalam mengolah data mentah yang berikutnya akan dimasukan ke dalam sistem klasifikasi [14]. Proses ini bertujuan untuk meningkatkan kualitas data, meningkatkan efisiensi dalam proses penambangan data [15]. Tahapan dalam *preprocessing text* adalah sebagai berikut:

a. Case Folding

Case Folding merupakan proses untuk mengubah semua karakter pada dokumen teks menjadi huruf kecil. Karakter yang diproses hanya huruf ‘a’ hingga ‘z’ dan selain itu juga karakter tersebut akan dihilangkan seperti tanda baca titik (.), koma (,), dan angka. [16] .

b. Tokenizing

Tokenization adalah proses memecah teks yang berupa kalimat, paragraf atau dokumen menjadi kata, frasa, simbol, atau elemen yang bermakna lain yang disebut sebagai token. Di dalam data dokumen, proses ini cukup sederhana karena spasi memisahkan kata-kata. [17]

c. Stopword

Stopword adalah kosakata yang bukan kata unik atau atribut karakter pada dokumen atau tidak memiliki informasi yang berarti pada teks atau kalimat. Yang dimaksud Kosakata pada pengertian *stopword* ialah kata penghubung dan kata keterangan yang tidak menjadi kata unik seperti “sebuah”, “karena”, “pada”,”ke” dll. [18].

d. Stemming

Stemming adalah proses yang bertujuan memperoleh kata dasar dari suatu token dengan cara menghilangkan awalan, akhiran, sisipan, dan confixes (kombinasi dari awalan dan akhiran).

2.3. Feature Extraction

Feature Extraction merupakan tahapan untuk mengubah teks atau dokumen yang sebelumnya masih berbentuk kata kedalam bentuk yang lebih representative, salah satunya *vector* [19]. Tujuannya agar teks atau dokumen dapat lebih mudah diklasifikasikan ke dalam kategori masing-masing. Pada tugas akhir ini digunakan metode ekstraksi fitur *weighting* TF-IDF. *TF (Term Frequency)* merupakan pembobotan kata (*term*) yang didasari pada perhitungan jumlah kata yang ada pada seluruh teks dokumen. *IDF (Inverse Document Frequency)* merupakan tingkat kepentingan suatu *term* dalam kumpulan dokumen atau pengukuran keunikan/karakter suatu *term* dalam suatu dokumen yang dibandingkan dengan dokumen lain [20]. Tujuan penggunaan *weighting*

(pembobotan) supaya nilai yang dihasilkan dari *feature extraction* tidak terlalu besar. Berikut adalah rumus dari *weighting* TF-IDF,

$$W_{i,d} = \{1 + log_{10}(tf_{i,d})\} \times \{log_{10}(N/df_i)\}$$

dengan keterangan $W_{i,d}$ adalah nilai pembobotan suatu *term* pada suatu dokumen, tf_i adalah jumlah kemunculan *term-i* pada dokumen-*d*, N merupakan jumlah keseluruhan dokumen, dan df_i merupakan jumlah dokumen yang terdapat *term-i*.

2.4. K-Nearest Neighbors

K-Nearest Neighbor adalah salah satu metode yang dipakai dalam klasifikasi data. Prinsip kerja *K-Nearest Neighbor* (K-NN) ialah melakukan klasifikasi data berdasarkan kedekatan jarak suatu data dengan data lainnya [21].

k-Nearest Neighbor
Classify (**X**, **Y**, *x*) // **X**: training data, **Y**: class labels of **X**, *x*: unknown sample
for *i* = 1 **to** *m* **do**
  Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set *I* containing indices for the *k* smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i$ where $i \in I\}$

Gambar 1 *Pseudocode k-Nearest Neighbor* [22]

Dekat atau jauhnya jarak bisa dihitung dengan besaran jarak. Dalam penelitian Tugas Akhir ini menggunakan Jarak *Euclidean*. Berikut adalah persamaan Jarak Euclidean :

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{p} (x_{ip} - x_{ij})^2}$$

dimana $d(x_i, x_j)$ adalah jarak Euclidean, x_{ip} adalah data *testing* ke-*i* pada variable ke-*p*, x_{ij} data *testing* ke-j pada variable ke-*p* dan *p* adalah dimensi data variable bebas.

2.5. Pengukuran Performa

Pengukuran performa dilakukan agar memperoleh tingkat kesesuain dari hasil klasifikasi dari sistem yang dibangun. Beberapa cara yang sering digunakan dalam pengukuran performa adalah dengan menghitung akurasi total, *recall*, *precission* dan *f1 measure* [23], dengan menggunakan persamaan sebagai berikut:

$$Precission_i = \frac{tp_i}{tp_i + fp_i}$$

$$Recall_i = \frac{tp_i}{tp_i + fn_i}$$

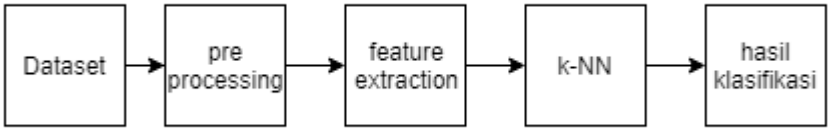
$$F1 - Score_i = 2 \times \frac{Precission_i \times Recall_i}{Precission_i + Recall_i}$$

dimana $Precission_i$ adalah presentase dari nilai item yang diprediksi benar dan terbukti benar [24, 20]. Nilai dari *precision* digunakan untuk mengukur seberapa tepat sistem melakukan prediksi , $Recall_i$ adalah presentase dari nilai item yang memang benar dan berhasil diprediksi benar [24] [20]. Nilai *recall* digunakan untuk mengukur seberapa banyak item yang memang diprediksikan benar , $F1 - Score_i$ adalah presentase dari nilai item yang merupakan kombinasi dari *Precison dan Recall* [24] [20].

3. Sistem yang Dibangun

3.1 Gambaran Umum Sistem

Gambaran umum dari sistem yang dibangun dimulai dengan tahap pengumpulan dataset, kemudian *preprocessing*, proses *feature extraction* menggunakan *weigthing TF-IDF*, proses klasifikasi menggunakan *k-NN*, dan hasil klasifikasi berita berupa nilai *micro average f1-measure*.



Gambar 2 Gambaran Umum Sistem

3.2 Dataset Berita

Data yang digunakan pada penelitian Tugas Akhir ini berupa teks berita yang terdiri dari bermacam-macam label kelas, seperti politik, budaya, ekonomi, otomotif, dll. Data teks berita diambil dimulai dari Agustus 2016 - Februari 2017, yang didapatkan dari berbagai portal berita yaitu, *website kompas.com*, *tribunnews.com*, *republika.com*, *mediaindonesia.com* dan *sindonews.com*. Dataset ini digunakan untuk membangun sistem klasifikasi berita menggunakan *k Nearest Neighbor* berdasarkan topik/kategori berita. Artikel berita bahasa Indonesia yang diambil dibagi menjadi 12 kelas yaitu: budaya, ekonomi, entertainment, hukum dan kriminal, kesehatan, lifestyle, otomotif, pendidikan, politik, sport, tekno dan wisata. Total jumlah record yang digunakan untuk pengklasifikasian artikel berita bahasa Indonesia yaitu 360 record.

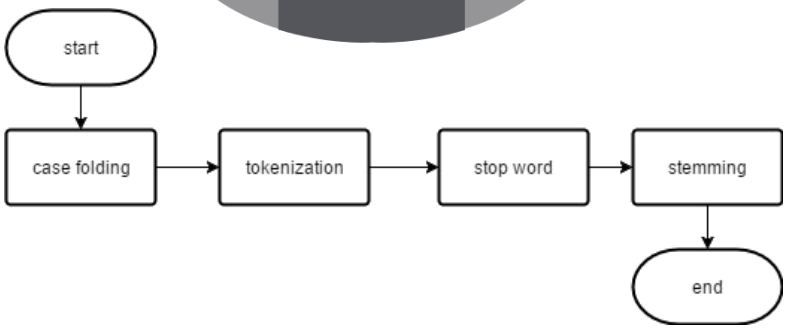
Tabel 1 Komposisi Data

Id Kelas	Nama Kelas	Jumlah Record	Id Kelas	Nama Kelas	Jumlah Record
1	Budaya	30	7	Otomotif	30
2	Ekonomi	30	8	Pendidikan	30
3	Entertainment	30	9	Politik	30
4	Hukum & Kriminal	30	10	Sport	30
5	Kesehatan	30	11	Tekno	30
6	Lifestyle	30	12	Wisata	30

Pada Tugas Akhir ini digunakan 12 kelas dengan komposisi seperti yang ditujukan pada table 1, Hal ini dikarenakan dari 5 media online yang digunakan oleh penulis terdapat 12 irisan kelas yang tersedia pada media online tersebut. Dikarenakan jumlah record yang digunakan untuk masing-masing kelas sebanyak 30 record, hal ini didukung dengan teori Central Limit pada buku *Introductory Statistics* milik Sheldon M Ross menyatakan bahwa aturan umum penggunaan ukuran sampel paling sedikit adalah sebanyak 30 [25].

3.3 Preprocessing

Preprocessing merupakan tahap awal untuk menyiapkan dataset untuk mempermudah d pemrosesan data, dan juga untuk mendapatkan tingkat performa yang tinggi. Dalam tahapan *preprocessing* didalamnya terdapat beberapa sub proses. Hai ini ditunjukkan pada gambar.



Gambar 3 Tahapan *preprocessing* dataset

3.4 Splitting Dataset

Tujuan dari *splitting dataset* adalah untuk mendapatkan data training dan data testing. Tentunya dalam menentukan data training dan data testing, diperlukan porsi data untuk pembagiannya. Akan

digunakan 3 skenario pembagian antara data *training* berbanding dengan data *testing* yaitu, 80%:20%, 50%:50%, 20%:80%. Dari 3 skenario pembagian data bertujuan untuk mendapatkan nilai performa sistem jika jumlah data *training* lebih banyak dari data *testing*, jumlah data *training* sama dengan jumlah data *testing* dan jika jumlah data *training* lebih sedikit dibandingkan dengan data *testing*.

3.5 Feature Extraction

Pada proses klasifikasi berita ini kata harus diubah menjadi bentuk vektor dengan menggunakan ekstraksi ciri TF-IDF.Pada tahap ini bobot dari setiap kata dihitung dengan jumlah kemunculan kata pada suatu dokumen dan jumlah dokumen yang mengandung kata tersebut.

Tabel 2 menunjukkan contoh jumlah kemunculan kata pada 6 dokumen. Dari tabel tersebut kita akan mencoba menghitung nilai dari TF-IDFnya. Untuk mempermudah perhitungan pertama kita akan

Tabel 2 Contoh Jumlah Kemunculan kata pada 6 dokumen

Kata	Jumlah Kemunculan					
	Dok_1	Dok_2	Dok_3	Dok_4	Dok_5	Dok_6
Musik	6	0	0	0	0	0
Daerah	4	0	3	1	1	0
Uang	1	2	1	0	0	1
Program	0	0	2	0	3	1
Hitung	1	4	0	2	1	0
Alat	5	1	1	0	0	2

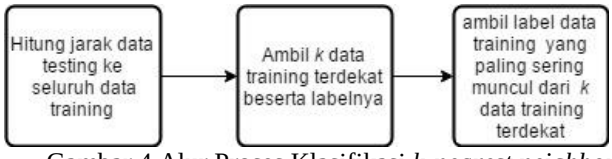
mencari nilai *tf* yaitu jumlah kemunculan kata-*i* dalam dokumen-*d* tersebut dan idf_i yaitu jumlah dokumen dibagi dengan jumlah dokumen keseluruhan yang memiliki kata-*i* yang bersangkutan. Misal diambil contoh perhitungan *weighting* TF-IDF untuk dokumen dok_1. Hasilnya ditampilkan pada Tabel 3 di bawah ini.

Tabel 3 Hasil *weigthing TF-IDF* Budaya_1

Kata	$1+\log(tf_{i,d})$	$\log(idf_i)$	$W_{i,d}=(1+\log(tf_{i,d}) \times \log(idf_i))$
Musik	2,7918	0,7782	2,1725
Daerah	2,3863	0,1761	0,4202
Uang	1	0,0969	0,0969
Program	0	0,3010	0
Hitung	1	0,3010	0,3010
Alat	2,6094	0,3522	0,9190

3.6 K-Nearest Neighbors

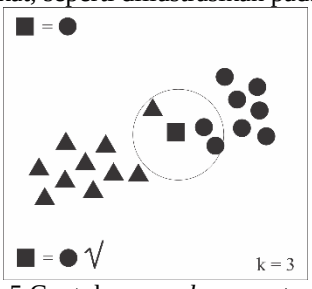
K-Nearest Neighbor merupakan salah satu metode yang digunakan dalam klasifikasi data. Prinsip kerja *K-Nearest Neighbor* (K-NN) adalah melakukan klasifikasi data berdasarkan kedekatan jarak suatu data dengan data yang lain [26]. Dekat atau jauhnya jarak bisa dihitung menggunakan besaran jarak. Dalam penerapannya seringkali digunakan jarak Euclidean. Jarak Euclidean adalah besarnya jarak suatu garis lurus yang menghubungkan antar objek.



Gambar 4 Alur Proses Klasifikasi *k-nearest neighbor*

Setiap Data *Testing* akan dihitung jaraknya terhadap seluruh Data *Training*. Perhitungan jarak tersebut dilakukan menggunakan persamaan *Euclidean Distance*. Setelah diperoleh jarak untuk seluruh Data *Training*, diambil sebanyak *k* Data *Training* terdekat dengan label data-data tersebut. Label yang paling sering muncul dari *k* Data *Training* tersebut akan menjadi label Data *Testing* yang

sedang diproses. Jika terdapat banyak kemunculan label data *Training* yang bernilai sama, maka akan diambil label data *Training* terdekat, seperti diilustrasikan pada gambar 5.



Gambar 5 Contoh proses *k-nearest neighbor* [15]

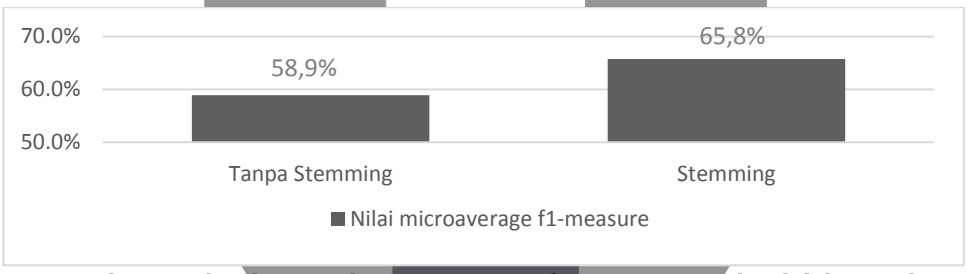
3.7 Evaluasi Sistem

Tahapan terakhir dalam perangcangan sistem klasifikasi topik berita adalah evaluasi sistem. Hasil proses tersebut ditunjukan pada gambar 3 proses klasifikasi dengan *k-nearest neighbors* digunakan untuk mengetahui performa sistem yang telah dibangun. Pengukuran performa dilakukan dengan menghitung nilai *precission, recall* dan *f-1 measure* yang sebelumnya telah dijelaskan sebelumnya di bab 2. Dalam pengukuran nilai *f1-measure (f1-score)* penulis menggunakan *micro average f1-measure*.

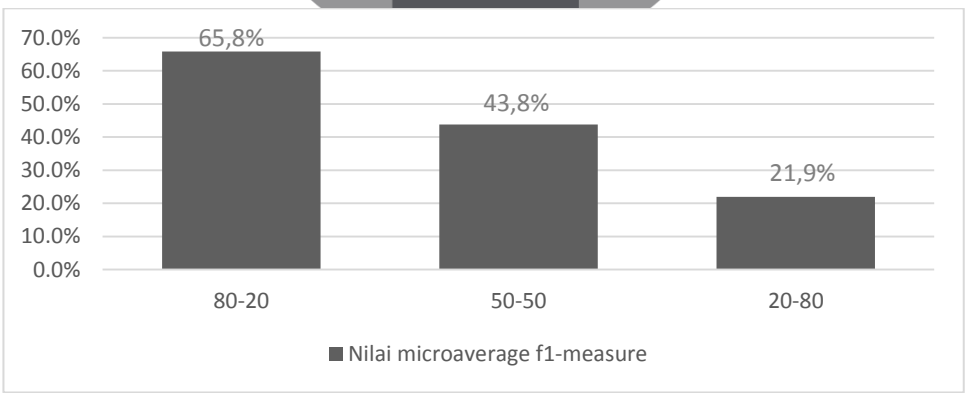
4. Evaluasi

4.1 Hasil Pengujian

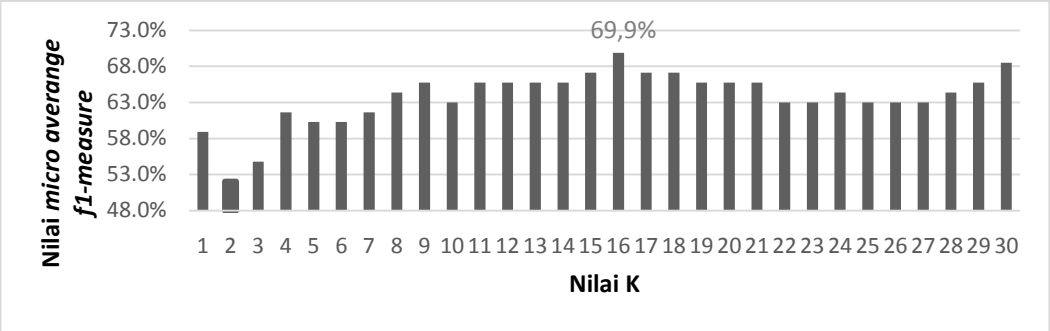
Hasil pengujian pada penelitian Tugas Akhir yang menggunakan 3 skenario pengujian ditunjukkan pada gambar 6, gambar 7 dan gambar 8. Gambar 6 menunjukkan skenario pengujian pertama dimana dilakukan dengan menggunakan dataset dengan proses stemming dan dataset tanpa proses stemming, kemudian pada gambar 7 merupakan hasil pengujian perbandingan nilai *micro average f1-measure* dengan skenario *splitting dataset.* Dan pada gambar 8 merupakan hasil pengujian perbandingan nilai *micro average* f1-*measure* pada nilai k= 1 sampai k=30.



Gambar 6 Perbandingan Nilai *micro average f1-measure* yang diambil dengan dan tanpa proses *Stemming*



Gambar 7 Perbandingan Nilai *micro average f1-measure* dengan skenario splitting dataset



Gambar 8 Perbandingan nilai *micro average f1-measure* pada tiap nilai K

4.2 Analisi Hasil Pengujian

Berdasarkan gambar 6 dapat dilihat bahwa dengan proses stemming pada tahap *preprocessing* mampu meningkatkan performa sistem yang dibangun sebesar 6.9 % atau mendapatkan nilai performa tertinggi dengan nilai *micro average f1-measure* sebesar 65,8%. Pada skenario pengujian tersebut menggunakan skenario *splitting dataset 80:20* dan menggunakan nilai *k=9*. Hal ini menunjukkan bahwa penggunaan *stemming* dapat meningkatkan performa sistem yang dibangun karena dengan proses *stemming* yang digunakan untuk mendapatkan kata dasar dari setiap kata pada berita bisa meningkatkan karakteristik dari setiap kelas berita

Kemudian pada gambar 7 menunjukkan perbandingan nilai *microaverage f1-measure* dengan skenario *splitting dataset*. Dari hasil tersebut menunjukkan bahwa jumlah data *training* mempengaruhi tingkat performa sistem, terlihat pada skenario *splitting dataset 80%-20%* dengan nilai *k=9* mendapatkan nilai *micro average f1-measure* tertinggi dengan nilai 65,8% dan pada *splitting dataset 20%:80%* mendapatkan nilai *micro average f1-measure* terendah. Hal ini disebabkan karena semakin banyak data *training* akan menambah informasi dari karakteristik setiap kelas berita.

Pada gambar 8 merupakan hasil dari perbandingan nilai *micro average f1-measure* pada nilai *k=1* sampai dengan *k=30*. Dari hasil skenario pengujian tersebut yang mendapatkan nilai *micro average* tertinggi adalah *k=16* dengan nilai *micro average f1-measure* sebesar 69,9%. %. Hal ini menunjukkan bahwa semakin besar nilai *k* juga mempengaruhi presentase tingkat performa dari sistem klasifikasi, kemudian jika nilai *k* sudah berada pada titik optimum yang mana pada system ini berada pada nilai *k* = 16 maka presentase tingkat performa system menurun. Hal ini disebabkan oleh semakin bertambahnya nilai k semakin bertambah juga data yang tidak memiliki kesamaan karakteristik terhadap data uji.

5. Kesimpulan

Dari hasil pengujian dan analisis yang dilakukan pada bab pengujian dan analisis dapat disimpulkan bahwa sistem yang dibangun dengan proses *stemming* dapat menambah performa sistem sebesar 6,9% kemudian rasio perbandingan data latih dan uji 80%:20% memiliki performa paling tinggi, sehingga dapat disimpulkan bahwa jumlah data latih yang semakin banyak akan menambah ketepatan klasifikasi dikarenakan sistem akan banyak mendapatkan informasi dari data latih. Pada penelitian ini jumlah *k* yang ditentukan pada *k-nearest neighbor* sangat berpengaruh. Dari sistem yang dibangun mendapatkan nilai *k =16* yang mendapatkan tingkat performa paling tinggi. Berdasarkan data yang sudah didapatkan, juga diketahui bahwa semakin besar nilai *k* pada awalnya akan memperbesar presentase ketepatan pada proses klasifikasi, kemudian jika nilai *k* sudah pada titik optimum maka besar presentase tingkat performa yang dibangun akan cenderung turun.

Adapun saran dari peneliti untuk penelitian selanjutnya yaitu Menambahkan jumlah dataset yang digunakan, karena dengan penambahan jumlah dataset diharapkan mampu menambahkan jumlah dan keragaman informasi sehingga dapat menambah informasi dari metode *k-nearest neighbors* dan meningkatkan performa sistem yang dibangun. Kemudian saran berikutnya adalah membuat sistem klasifikasi topik berita bahasa Indonesia yang mampu menangani masalah *multilabel classification*, karena pada suatu berita terkadang memiliki kategori yang banyak dikarenakan informasi yang tertera didalam berita sangat beragam, sebagai contoh suatu berita mengandung informasi tentang politik dan hukum, atau suatu berita dapat juga mengandung informasi tentang teknologi, ekonomi dan gaya hidup dalam waktu yang bersamaan.

Daftar Pustaka

[1] KBBI, "KBBI," 2016. [Online]. Available: http://kbbi.web.id/berita/. [Accessed 9 April 2017].

[2] Okezone, "Okezone," 16 Maret 2016. [Online]. Available: http://economy.okezone.com/read/2016/03/16/320/1337230/96-masyarakat-indonesia-konsumsi-berita-online. [Accessed 9 April 2017].

[3] M. Harlian, Machine Learning Text, Austin: University of Texas, 2006.

[4] A. P. Jain and V. D. Katkar, "Sentiment Analysis of Twitter data using data mining," *International Conference on Information Processing (ICIP),* pp. 807 - 810, 2015.

[5] A. D. Saputra, Adiwijaya and M. S. Mubarok, "Klasifikasi Sentimen Pada Level Aspek Terhadap Ulasan Produk Berbahasa Inggris Menggunakan Bayesian Network (case Study: Data Ulasan Produk Amazon)," *eProceedings of Engineering*, vol. 4, no. 3, 2017.

[6] R. A. Aziz, M. S. Mubarok and Adiwijaya, "Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes," *Indonesia Symposium on Computing (IndoSC)*, 2016.

[7] M. S. Mubarok, Adiwijaya and M. D. Aldhi, "Aspect based sentiment analysis to review products using naive bayes," *AIP Conference Proceeding 1867*, vol. 1867, p. 020060.

[8] M. H. Syahnur, M. S. Mubarok and M. A. Bijaksana,"Kategorisasi topik tweet di kota Jakarta, Bandung, dan Makassar dengan metode Multinomial Naïve Bayes classifier," *eProceeding of Engineering* , vol. 3, no. 2.

[9] M. A. Nurjaman, M. S. Mubarok and Adiwijaya, "Analisis Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Support Vector Machine," *eProceeding of Engineering*, vol. 4, no. 3.

[10] L. R. Putri, M. S. Mubarok and Adiwijaya, "Klasifikasi Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Naïve Bayes," *eProceeding of Engineering*, vol. 4, no. 3, 2017.

[11] Adiwijaya, M. N. Aulia, M. S. Mubarok, U. N. Wisesty and F. Nhita, "A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system," *Information and Communication Technology (ICoIC7), 2017 5th International Conference on*, pp. 1-5, 2017.

[12] R. Ahyan, W. N. Sikumbang and S. R. Henim, "Implementasi Text Mining dengan K-Nearest Neighbors Pada Analisis Sentimen," 2016.

[13] N. Supraptiningsih, M. S. Mubarok and Adiwijaya, "Analisis sentimen level aspek pada ulasan produk menggunakan k-nearest neighbor," 2017.

[14] J. Han, M. Kamber and J. Pei, Data Mining Concepts and techniques, San Fransisco: Morgan Kauffman , 2012.

[15] Suyanto, Data Mining, Bandung: Penerbit Informatika, 2017.

[16] S. Weiss, Text Mining: Predictive Methods of Analyzing Unstructured Information, New York: Springer, 2010.

[17] M. P.O and R. D.L, Data Mining and Knowledge Discovery, New York : Springer, 2010.

[18] E. Dragut, F. Fang, P. Sistla, C. Yu and W. Meng, Stop Word and Related Problems, Proc: VLDB Endowment, 2009.

[19] C. S. Utami, M. A. Mukid and Sugito, "Klasifikasi Kinerja Perusahaan di Indonesia dengan menggunakan Metode Weigthed K Nearest Neighbor (Studi Kasus : 436 Perusahaan yang terdaftar di Bursa Efek Indonesia Tahun 2015)," 2017.

[20] R. A. Sasanty, M. S. Mubarok and Adiwijaya, "Analisis Sentimen Level Aspek Pada Ulasan Produk Menggunakan Support Vector Machine," 2017.

[21] E. Prasetyo, Data Mining Konsep, Yogyakarta: ANDI Yogyakarta, 2012.

[22] B. Tay, J. Hyuan and S. Oh, "A machine learning approach for specification of spinal cord injuries using fractional anisotropy values obtained from diffusion tensor images," *Computational and mathematical methods in medicine*, 2014.

[23] A. Hotho, A. Nurnberger and G. Paass, A Brief Survey of Text, Kassel: University of Kassel, 2005.

[24] Sokolova, Marina and G. Lapalme , "Information Processing and Management," doi:10.1016/j.ipm.2009.03.002.

[25] S. M. Ross, Introduction to Probability and Statistic For Engineers and Scientist, Fifth Edition, 2014.

[26] E. Prasetyo, Data Mining Konsep dan Aplikasi mengunakan MATLAB, Yogyakarta: ANDI Yogyakarta, 2012.

[27] Adiwijaya, Aplikasi Matriks dan Ruang Vektor, Yogyakarta: Graha Ilmu, 2014.

[28] Adiwijaya, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta, 2016.