Klasifikasi Topik Berita Berbahasa Indonesia menggunakan *Weighted K-Nearest Neighbor*

Sigit Bagus Setiawan[1], Adiwijaya[2], Mohamad Syahrul Mubarok[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung

[1]sbastian@student.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id, [3]msyahrulmubarok@gmail.com

Abstract

News is one of the means of information for the general public. In this modern era, many people use online media as one means to access news. In Indonesia online media has the largest percentage as a means of delivering the news [1]. But the number of news that exist in the online media raises the problem in categorizing news topics that exist. Therefore it takes a system to be able to categorize every news topic that is on online media. This study aims to create a system that is able to categorize every news in Indonesian on the class that should be. Classification uses the weighted k-Nearest Neighbor (wkNN) because this is a simple and powerful classifier. In this research there are several stages in system design, namely preprocessing data, feature extraction, and classification using weighted k-Nearest Neighbor. After the stages are passed the performance measurement is done. The result of research can give performance system equal to 75,86% with value of k=27.

Keywords: *classifier, preprocessing, feature extraction, weighted k-NN*

Abstrak

Berita adalah salah satu sarana informasi bagi masyarakat umum. Di jaman yang modern ini orang-orang banyak menggunakan media *online* sebagai salah satu sarana untuk mengakses berita. Di Indonesia sendiri media *online* memiliki presentase paling besar sebagai sarana penyampaian berita [1]. Namun banyaknya berita yang ada dalam media *online* memunculkan masalah dalam mengkategorikan topik berita yang ada. Sehingga dibutuhkanlah sistem yang dapat mengkategorikan setiap topik berita yang ada pada media *online*. Penelitian ini bertujuan menciptakan sistem yang mampu mengkategorikan setiap berita berbahasa Indonesia pada kelas yang seharusnya. Pengklasifikasian menggunakan metode *weighted k-Nearest Neighbor* (*wkNN*) karena merupakan *classifier* yang sederhana namun *powerful*. Pada penelitian ini terdapat beberapa tahap dalam perancangan sistem, yaitu *preprocessing* data, *feature extraction*, dan pengklasifikasian menggunakan *weighted k-Nearest Neighbor*. Setelah tahap-tahap tersebut dilalui dilakukanlah pengukuran performansi. Hasil penelitian mampu memberikan performa sistem sebesar 75,86% dengan nilai k = 27.

Keywords: *classifier, preprocessing, feature extraction, weighted k-NN*

I.ATAR BELAKANG

Berita merupakan salah satu sarana informasi yang memungkinkan kita untuk mendapatkan pengetahuan yang baru, sehingga tak dipungkiri berita merupakan salah satu sarana informasi yang penting bagi masyarakat umum [1]. Banyak media yang digunakan untuk menyajikan berita kepada masyarakat, seperti media cetak, elektronik, maupun online. Di Indonesia sendiri media online menjadi salah satu media favorit yang sering digunakan masyarakat untuk mendapatkan berita tentang kondisi terkini yang mereka butuhkan. Bahkan berdasarkan sebuah laporan dari UCWeb konsumsi berita online di Indonesia meningkat 61 persen dari tahun 2015 [2]. Dalam media online berita yang disajikan dibagi dalam beberapa klasifikasi topik berita untuk memudahkan masyarakat dalam mengakses berita dengan topik yang mereka inginkan. Dalam website online sendiri banyaknya artikel berita yang masuk menimbulkan masalah pelabelan topik berita untuk setiap artikel. Untuk itu dibutuhkanlah metode yang dapat mengklasifikasikan setiap topik berita dengan efisien dan

akurat untuk memudahkan penyedia layanan media online dalam pelabelan berita, sehingga masyarakatpun juga akan terbantu dalam pengaksesan berita.

Proses mengidentifikasi dokumen ke dalam kelas-kelas yang sebelumnya sudah terdefinisi inilah yang disebut proses klasifikasi. Dalam proses klasifikasi pemilihan metode yang digunakan menjadi hal yang berpengaruh dalam kinerja klasifikasi. Banyak metode yang bisa digunakan dalam pengklasifikasian topik berita berbahasa Indonesia, namun dalam tugas akhir ini diharapkan sebuah *classifier* yang sederhana namun *powerful* agar pelabelan bisa dilakukan secara mudah dan cepat. Salah satu metode yang sederhana dalam proses klasifikasi adalah *k-Nearest Neighbor* (k-NN).

Namun pada k-NN biasa bobot setiap data pada k data terdekat besarnya sama yaitu 1/k. Padahal jika bobotnya diatur (tidak 1/k) maka *classifier* dapat memiliki performansi yang lebih bagus [3, 16, 18]. Maka dari itu pada tugas akhir ini metode yang akan digunakan adalah pengembangan dari k-NN yaitu *weighted k-Nearest Neighbor* yang diharapkan dapat memberikan performansi yang lebih baik daripada metode k-NN biasa dalam klasifikasi teks berita berbahasa Indonesia.

II. TINJAUAN PUSTAKA

A. *Text Mining*

Text mining mempunyai pengertian menambang data teks yang mana sumber untuk data biasanya di dapatkan dari dokumen-dokumen yang bertujuan mencari kata yang bisa merepresentasikan isi dokumen-dokumen hingga bisa dilakukan analisis hubungan tiap dokumen-dokumen [4]. *Text mining* adalah proses ekstraksi pola berupa informasi dan pengetahuan yang dapat digunakan dari data teks dalam jumlah besar, yaitu dokumen Kutipan teks, PDF, Word, dan lain sebagainya. Sama seperti data *mining*, *text mining* juga memiliki tujuan untuk melakukan ekstraksi informasi dari data yang ada melalui identifikasi dan penelusuran pola yang menarik. Namun, dalam *text mining* sumber data yang digunakan berupa sumber data tekstual dimana tidak memiliki struktur yang formal seperti layaknya database yang terstruktur. *Text mining* bertujuan guna memperoleh informasi yang bernilai dari dokumen-dokumen yang ada. Data-data yang dipakai dalam *text mining* merupakan sekumpulan teks yang formatnya tidak terstruktur. *text mining* memiliki tugas khusus yaitu mengkategorikan teks dan mengelompokkan teks. *Text mining* adalah pengimplementasian konsep dan teknik dari data *mining* guna mencari suatu pola di dalam teks. Berdasar struktur data teks yang tidak teratur proses *text mining* membutuhkan tahap awal untuk persiapan supaya teks bisa diubah menjadi teks yang lebih terstruktur.

B. *Text Prepocessing*

Salah satu tantangan dari *text mining* yaitu mengubah teks yang tidak terstruktur dan semi terstruktur ke dalam model vektor ruang terstruktur [5]. Hal ini merupakah langkah yang harus dilakukan sebelum melakukan sesuatu terkait text mining lebih lanjut ataupun analisis. Beberapa preprocessing yang sering digunakan adalah: *Case folding* adalah proses konversi huruf besar menjadi huruf kecil. Selain itu juga pada tahap ini, karakter bukan huruf maka akan dieliminasi, misal seperti karakter (, ! ,-, ?, &, dll). *Tokenization* merupakan proses memecah kalimat atau string input berdasarkan setiap kata yang menyusun dokumen. *Stopword Removal* merupakan proses menghilangkan kata yang sering muncul namun tidak memiliki arti atau tidak relevan dalam *information rediscovery* [13]. Contoh *stopword* dalam bahasa Indonesia adalah “dan”, “ke”, “saya” , “yang” dll. Pada tugas akhir ini digunakan *stopword removal* untuk berita bahasa Indonesia dari hasil penelitian yang dilakukan oleh Liyantato [7, 14].

C. *Feature Extraction*

Feature Extraction merupakan tahapan untuk mengubah teks atau dokumen yang sebelumnya masih berbentuk kata kedalam bentuk vektor [15]. Tujuannya agar teks atau dokumen dapat diklasifikasikan ke dalam kelas-kelas yang sudah ada. Pada tugas akhir ini digunakan metode ekstraksi fitur *weighting* TF-IDF.

D. Euclidean Distance

Metode pengukuran jarak yang paling umum digunakan untuk menghitung *square distance* dari dua vector dan didefinisikan dengan menggunakan persamaan berikut [8]:

$$Dist_{xy} = \sqrt{\sum_{k=1}^{n} (x_{ik} - y_{jk})^2} \qquad (1)$$

x_{ik} adalah nilai objek ke satu (data latih) , y_{jk} adalah nilai objek ke dua (data uji) dengan n sebagai dimensi data , dan k yaitu variabel data.

III. DESAIN SISTEM

Metode penelitian dilakukan dengan mengimplementasikan *weighted* K-NN kedalam perangkat lunak Matlab. Rancangan sistem klasifikasi dibuat menjadi 3 tahap utama sesuai yang ditunjukkan oleh Gambar 2. Ekstraksi ciri pada penelitian ini berdasarkan *weighting* TF-IDF.

Gambar 3: Rancangan sistem Klasifikasi menggunakan K-NN

A. Preprocessing

Preprocessing merupakan tahapan awal yang dilakukan untuk mengolah data yang belum sesuai dengan bentuk data yang diharapkan, seperti: case folding, tokenization, stopwords removal, dan stemming dengan tujuan dilakukannya penyeragaman untuk mempermudah kegiatan pemrosesan data.

B. Ekstraksi Ciri

Pada proses klasifikasi berita ini kata harus diubah menjadi bentuk vektor dengan menggunakan ekstraksi cirri TF-IDF [15, 17]. Pada tahap ini bobot dari setiap kata dihitung dengan memperhatikan TF-IDF (*Inverse Document Frequency*) yang merupakan salah satu algoritma text mining yang paling terkenal yang digunakan dalam penelitian. Yang merupakan sebuah algoritma yang digunakan untuk menghitung probabilitas Invers dari mencari kata dalam teks. Berikut adalalah rumus dari TF-IDF.

$$W = 1 + \log(tf) \times log10(N/df)$$

Dengan keterangan W adalah hasil ekstraksi cirri TF-IDF, tf adalah jumlah kemunculan kata dalam dokumen, N merupakan jumlah keseluruhan dokumen, dan df merupakan jumlah dokumen-dokumen yang memiliki kata yang bersangkutan.

C. Klasifikasi Weighted K-NN

Hasil ekstraksi ciri adalah masukan informasi data untuk diklasifikasikan dengan K-NN. Gambaran secara umum pada K-NN, yaitu dihitung jarak dari data uji terhadap semua data latih. Kemudian diambilah sejumlah k pada data latih terhadap jarak yang terdekat. Data uji dapat ditentukan berdasarkan kategori mayoritas dari tetangga yang terdekat.

Algoritma *Weighted* K-NN [8]:

1. Mulai masukan data latih, dan data uji
2. Tentukan parameter K
3. Menghitung jarak objek ke semua data latih dengan metode *euclidean distance*.
4. Normalisasi jarak euclide dengan rumus *dbaru=d/k+1*
5. Urutkan hasil perhitungan jarak secara *ascending*

6. Pembobotan pada jarak dengan menggunakan kernel gauss atau *simply method*
7. Tentukan kelompok data uji berdasarkan kelas dengan bobot terbesar

Penelitian ini menggunakan parameter k=1-30 dengan metode *euclidean distance* yang dijadikan sebagai pengukuran jarak. Dalam pembagian data latih dan uji dilakukan *train test split* yaitu membuat perbandingan data sebesar 80:20, 50:50,dan 30:70 persen. Data latih dan uji yang digunakan bersifat acak. Hasil akurasi dapat dibandingkan berdasarkan *train or test split* tersebut.

D. Penelitian Terkait weighted k-NN

k-NN merupakan algoritma yang umum dan banyak diimplementasikan pada beberapa hal, namun masih sedikit pengembangan dari algoritma ini (*weighted* k-NN) yang dapat ditemukan dalam penelitian. Sedikit diantaranya adalah penelitian oleh Hechenbichler Schliep, *Weighted K-Nearest Neighbor Techniques and Ordinal Classification*. Dalam penelitian ini dijabarkan teknik klasifikasi dengan menggunakan *weighted k-Nearest Neighbor*. Secara garis besar ada lima langkah pengklasifikasian dengan menggunakan *weighted* k-NN, yang pertama adalah menentukan sampling data, selanjutnya mencari nilai k+1, dilanjutkan dengan normalisasi jarak *euclide*, lalu kemudian dilakukan pembobotan pada jarak yang sudah dinormalisasi dan mencari bobot terberat untuk menentukan kelas klasifikasi.

Pengaplikasian metode *weighted* k-NN juga dilakukan oleh Cyntia Surya Utami [15] dimana dilakukan pengklasifikasian kinerja perusahaan di Indonesia dengan menggunakan metode *weighted k-Nearest Neighbor* dengan studi kasus 436 perusahaan yang terdaftar di bursa efek Indonesia pada tahun 2015. Dalam penelitian ini teknik dan tahapan langkah yang digunakan sama dengan penelitian yang dilakukan oleh Hechenbichler Schliep. Pada hasil akhirnya didapatkan k optimal 3 dan akurasi sebesar 98,8%.

IV. HASIL PENGUJIAN DAN DISKUSI

Berdasarkan metode penelitian yang telah dirancang dan diujikan maka diperoleh hasil-hasil sebagai berikut ini:

A. Hasil Ekstraksi Ciri

Untuk mengetahui hasil ekstraksi ciri teks berita digunakan *weighting* TF-IDF dengan hasil bisa dilihat pada tabel 2 dari contoh perhitungan kata pada Tabel 1.

Kata	Jumlah Kemunculan					
	Budaya_1	Budaya_2	Ekonomi_1	Ekonomi_2	Tekno_1	Tekno_2
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

TABEL I. CONTOH JUMLAH KEMUNCULAN

Kata	Nilai *Weighting* TF-IDF					
	Budaya_1	Budaya_2	Ekonomi_1	Ekonomi_2	Tekno_1	Tekno_2
Musik	2,1725	0	0	0	0	0
Daerah	0,4202	0	0,3695	0,1761	0,1761	0

TABEL II. HASIL EKSTRAKSI CIRI SESUAI DENGAN TABEL I

Uang	0,0969	0,1633	0,0969	0	0	0,0969
Program	0	0	0,5096	0	0,6316	0,3010
Hitung	0,3010	0,7182	0	0,5096	0,3010	0
Alat	0,9190	0,3522	0,3522	0	0	0,5963

Setelah mengubah kata dalam bentuk vektor dengan mencari nilai weighting TF-IDF, proses selanjutnya adalah dengan menggunakan metode *weighted k-Nearest Neighbor*.

B. Confusion Matrix

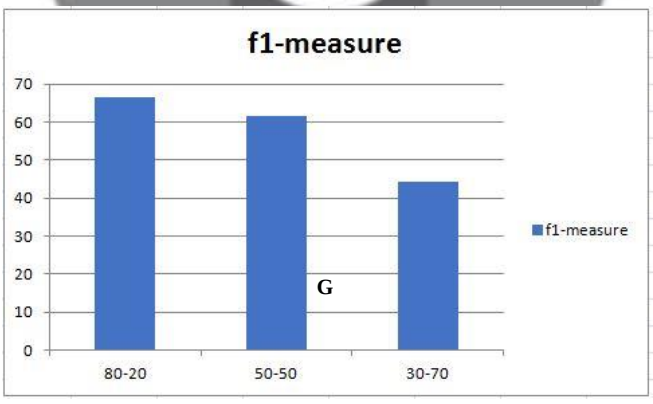
Hasil pengujian klasifikasi dapat dilakukan dengan confusion matrix untuk mengetahui kinerja sistem klasifikasi *weighted* k-NN. Pada *confusion matrix* terdapat 4 istilah sebagai representasi hasil proses klasifikasi yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN).

Berdasarkan empat istilah tersebut diperoleh hasil akurasi, sensitivitas, dan spesifisitas. Pada penelitian ini sensitivitas merupakan tingkat akurasi model dalam klasifikasi kelas *true* yaitu sebagai artikel yang berhasil dikategorikan dengan benar.

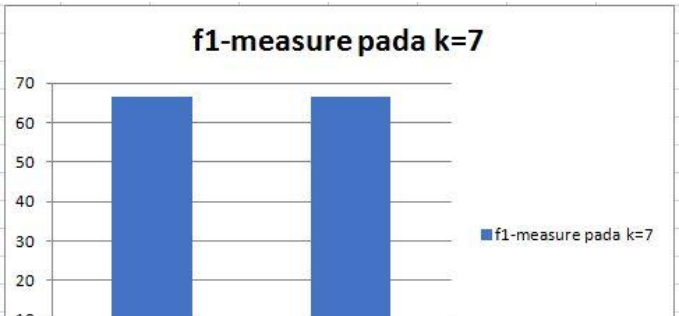
C. Hasil Pengujian Data

Dalam klasifikasi berita yang dilakukan dengan beberapa skenario pengujian data berdasarkan *train or test data split* menggunakan parameter nilai K=2-30. *Train or test data split* merupakan cara pembagian data keseluruhan menjadi data latih dan data uji.

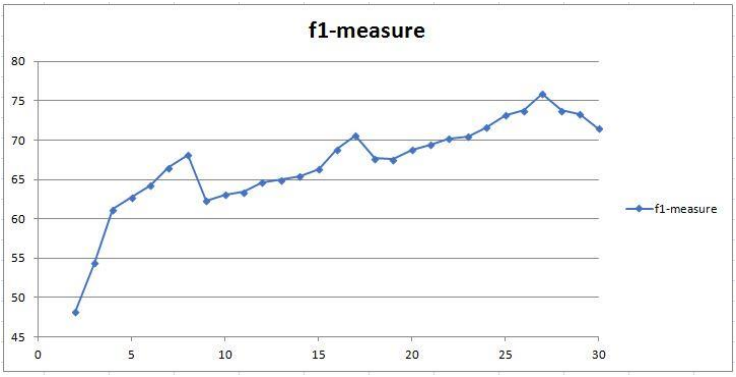
Pada penelitian ini menggunakan tiga skema pembagian data, yaitu data latih dan data uji masing-masing 80:20, 50:50, dan 30:70 persen secara acak. Pembagian data ini dilakukan agar dapat menguji kinerja metode penelitian untuk beberapa komposisi data. Sehingga, masing-masing skema memiliki tingkat akurasi yang berbeda dan dapat dibandingkan untuk mengetahui skema akurasi terbaik. Dibawah ini adalah gambar dari hasil pengujian data yang telah dilakukan



ambar 1: Perbandingan f1-measure pada data split



Gambar 2: Perbandingan f1-measure pada simply method dan kernel gauss



Gambar 3: Perbandingan f1-measure pada tiap k

D. Analisis Hasil Pengujian

Berdasarkan hasil pengujian dengan tiga skema pembagian dataset memperoleh akurasi yang berbeda-beda. Dari keseluruhan skema. Beberapa faktor yang mempengaruhi hasil sistem tersebut, yaitu :

a. Pembagian dataset

Pembagian dari data latih dan data uji yaitu dengan *train or test data split* menjadi tiga bagian skema. Nilai *f1-measure* dari skema 1,2, dan 3 berturut-turut yaitu 66,54%, 61,56%, dan 44,19%. Terlihat bahwa, semakin banyak data latih yang digunakan maka peluang nilai akurasi sistem semakin besar. Skema 1 dengan pembagian data uji dan data latih sebanyak 80:20 persen memiliki akurasi tertinggi ketika parameter k=7.

b. Parameter K (Jumlah Tetangga Terdekat)

Pada tabel diatas kita bisa tahu k mana yang paling optimal. Dari penentuan k yang dilakukan secara manual oleh penulis dapat diketahui bahwa k (tetangga terdekat) untuk mendapatkan performansi maksimum yaitu pada jumlah k = 27 dengan nilai *f1-measure* sebesar 75,86%. Berdasarkan data yang sudah didapatkan kita juga mengetahui bahwa semakin besar nilai k pada awalnya akan memperbesar presentase kebenaran pada klasifikasi, kemudian jika k sudah pada titik optimum maka besar presentase akan cenderung turun. Hal tersebut sesuai dengan penelitian Mutiara Ayu Banjarsari [18] yang meneliti tentang Penerapan K-Optimal Pada Algoritma k-NN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4.

E. Rekomendasi Sistem

Berdasarkan hasil analisis pengujian maka saran untuk pengembangan sistem selanjutnya dalam klasifikasi teks berita berbahasa Indonesia, yaitu :

1. Menambahkan jumlah dataset untuk data sampel, karena dengan dataset yang lebih banyak diharapkan dapat menambah keragaman dan jumlah informasi sehingga dapat menambah referensi dari metode *weighted* k-NN dan meningkatkan performa dari sistem.

2. Membuat sistem yang mampu mengklasifikasikan artikel berita bahasa Indonesia dalam penanganan masalah *multilabel*, karena ada berita yang terkadang didalamnya mengandung informasi yang berisi lebih dari satu label kelas atau kategori. Contohnya ketika sebuah berita berisi tentang informasi budaya dan teknologi sekaligus.

V. KESIMPULAN

Dari hasil pengujian dan analisis yang telah dilakukan pada sebelumnya dapat disimpulkan bahwa:

1. Sistem yang dibangun dengan perbandingan data training dan testing 80%:20% memiliki performa terbaik. Sehingga dapat disimpulkan bahwa jumlah data training yang semakin banyak akan menambah ketepatan klasifikasi dikarenakan sistem akan banyak mendapatkan informasi dari data training.
2. Pada penelitian ini dapat disimpulkan bahwa pemilihan pemberian bobot pada weighted k nearest neighbor cukup berpengaruh atas performansi sistem.

3. Pada penelitian ini jumlah k yang ditentukan pada klasifier weighted k nearest neighbor sangat berpengaruh. Dari penentuan k yang dilakukan secara manual oleh penulis dapat diketahui bahwa k (tetangga terdekat) untuk mendapatkan performansi maksimum yaitu pada jumlah k = 27 dengan performansi sebesar 75,86%. Berdasarkan data yang sudah didapatkan kita juga mengetahui bahwa semakin besar nilai k pada awalnya akan memperbesar presentase kebenaran pada klasifikasi, kemudian jika k sudah pada titik optimum maka besar presentase akan cenderung turun.

.

DAFTAR PUSTAKA

[1] Wikipedia, "Berita," 2016. [Online]. Available: https://id.wikipedia.org/wiki/Berita [Accessed 7 April 2017].
[2] Technologue, "Konsumsi berita online naik 61 Indonesia darurat berita," 2017. [Online]. Available: http://technologue.id/konsumsi-berita-online-naik-61-indonesia-darurat-berita/ [Accessed 7 April 2017].
[3] Stone C. J, Consistent Nonparametric Regression, 1977.
[4] H. Witten Ian, *Text mining* and Analytics, 2013.
[5] Famili Fazel, Min Shen Wei, Weber Richard, Simoudis Evangelos, Data Preprocessing and Intelligent Data Analysis, 1997
[6] M. P.O and R. D.L, Data Mining and Knowledge Discovery, New York: Springer, 2010.
[7] E. Dragut, F. Fang, P. Sistla, C. Yu and W. Meng, Stop Word and Related Problems, Proc: VLDB Endowment, 2009.
[8] H. Klaus, S. Klaus, Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, 2004.
[9] Wordpress, "Perbedaan *precision*, *recall*, *fi-measure*," 2013. [Online].Available at: https://dataq.wordpress.com/2013/06/16/perbedaan-precision-recall-accuracy/ [Accessed 7 April 2017].
[10] Jumadi dan W. Edi, Penggunaan KNN (k-Nearest Neighbor) untuk Klasifikasi Teks Berita yang tak Terkelompokkan pada saat Pengklasteran oleh STC(Suffic Tree Clustering), 2015.
[11] K. Bambang, E. Syahril, dan S. Salim, Klasifikasi Konten Berita Dengan Metode Text Mining, 2012.
[12] A. Risyad, W. N. Sikumbang dan S. R. Henim, Implementasi Text Mining dengan K-Nearest Neighbor pada Analisis Sentimen, 2016.
[13] Wordpress, "Kata Dasar Bahasa Indonesia," 2010. [Online]. Available at: http://liyantanto.wordpress.com/2010/12/06/kata-dasar-bahasa-indonesia/ [Accessed 22 November 2017]
[14] Stanford, "Ranked Information Retrieval," 2012. [Online]. Available at: https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html [Accessed 22 November 2017]
[15] C. S. Utami, M. A. Mukid, Sugito, Klasifikasi Kinerja Perusahaan di Indonesia dengan Menggunakan Metode Weighted K Nearest Neighbor (Studi Kasus: 436 Perusahaan Yang Terdaftar Di Bursa Efek Indonesia Tahun 2015), 2017
[16] Adiwijaya, Aplikasi Matriks dan Ruang Vektor, Yogyakarta: Graha Ilmu, 2014.
[17] Adiwijaya, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta, 2016.
[18] M. A. Banjarsari, Penerapan K-Optimal Pada Algoritma k-NN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4, 2017