PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DENGAN DECISION RULE UNTUK KLASIFIKASI SUBTOPIK BERITA

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 16 Januari 2014

YOSEPH SAMUEL
22094664

HALAMAN PERSETUJUAN

Judul Skripsi	:	IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DENGAN DECISION RULE UNTUK KLASIFIKASI SUBTOPIK BERITA
Nama Mahasiswa	:	YOSEPH SAMUEL
N I M	:	22094664
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2013/2014

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 20 Desember 2013

Dosen Pembimbing I	Dosen Pembimbing II
Rosa Delima, S.Kom., M.Kom.	Antonius Rachmat C., SKom.,M.Cs

Lembar Pengesahan

Skripsi dengan judul:

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DENGAN DECISION RULE UNTUK KLASIFIKASI SUBTOPIK BERITA

telah diajukan dan dipertahankan oleh:

YOSEPH SAMUEL
22094664

dalam Ujian Skripsi Program Studi Teknik Informatika
Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
dan dinyatakan DITERIMA untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer pada tanggal 9 Januari 2014

Nama Dosen	Tanda Tangan
1. Rosa Delima, S.Kom., M.Kom.	
2. Antonius Rachmat C., SKom.,M.Cs	
3. Aditya Wikan Mahastama, S.Kom.	
4. Yuan Lukito, S.Kom., M.Cs.	

Yogyakarta,
Disahkan Oleh:

Dekan,

(Drs. Wimmie Handiwidjojo. MIT.)

Ketua Program Studi,

(Nugroho Agus Haryono. M.Si)

UCAPAN TERIMA KASIH

Puji dan syukur kehadirat Tuhan Yang Maha Esa, yang telah melimpahkan rahmat dan karunia kepada penulis dapat menyelesaikan Skripsi dengan judul Implementasi Algoritma *K-Nearest Neighbor* dengan *Decision Rule* untuk Klasifikasi Subtopik Berita.

Penulisan dan penyusunan Skripsi ini disusun dalam rangka melengkapi syarat untuk memperoleh gelar Sarjana Komputer. Selain itu bertujuan melatih mahasiswa untuk menghasilkan suatu karya yang dapat dipertanggungjawabkan secara ilmiah, dan dapat bermanfaat bagi penggunanya.

Pada kesempatan ini, penulis mengucapkan terima kasih kepada semua pihak yang telah membantu dalam menyusun skripsi, antara lain :

1. **Ibu Rosa Delima, S.Kom, M.Kom** selaku dosen pembimbing I yang telah memberikan bimbingannya serta memberi masukkan yang sangat membantu dari awal hingga akhir selesainya Skripsi ini, juga kepada
2. **Bapak Antonius Rachmat, S.Kom, M.Cs.** selaku dosen pembimbing II yang memberikan petunjuk dan masukkan dari awal hingga akhir selesainya Skripsi ini.
3. Kedua Orang Tua tercinta yang selalu memberikan semangat, perhatian, dan motivasi serta dukungan agar Skripsi ini selesai.
4. Yessy, Henry, Albert, Yuda, Dimas dan pihak lain yang tidak dapat penulis sebut satu-persatu yang telah memberikan semangat dan masukan, sehingga Skripsi ini dapat terselesaikan.

Akhir kata, dengan kerendahan hati, penulis menyadari bahwa Skripsi ini masih jauh dari sempurna, oleh karena itu penulis menerima kritik, saran, semoga Skripsi ini dapat bermanfaat bagi semua pihak.

Yogyakarta, Desember 2013

Yoseph Samuel

KATA PENGANTAR

Puji syukur penulis ucapkan kepada Tuhan Yesus atas berkat dan penyertaannya sehingga penulis dapat menyelesaikan Skripsi yang berjudul **Implementasi Algoritma K-Nearest Neighbor dengan Decision Rule untuk Subtopik Berita.**

Membaca berita melalui internet sudah menjadi kebiasaan setiap masyarakat pada jaman era digital ini. Tiap masyarakat mempunyai ketertarikan pada topik berita yang berlainan. Penggunaan internet sebagai sarana untuk membaca berita hampir menggantikan kegunaan koran, majalah dan beberapa sumber berita fisik lainnya. Pengklasifikasian dokumen merupakan salah satu cara untuk mengklasifikasi tiap topik yang masyarakat inginkan. Karena banyaknya berita yang beredar di internet, maka klasifikasi berita diperlukan agar pembaca lebih mudah mencari berita yang diinginkan.Sistem yang penulis rancang menggunakan algoritma *K-Nearest Neighbor*. Dengan menggunakan *euclidean distance* sebagai *Decision Rule*, penulis ingin merancang sistem yang dapat mengklasifikasi subtopik dari topik berita.

Melalui penulisan Skripsi ini, penulis berharap agar metode dalam pengenalan karakter dapat semakin dikembangkan. Berbagai metode maupun algoritma yang berbeda dapat digunakan agar pengenalan karakter semakin baik. Tidak menutup kemungkinan juga akan ditemukannya metode baru dalam pengenalan karakter ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, maka saran dan kritik yang konstruktif dari semua pihak sangat diharapkan demi penyempurnaan selanjutnya.

Yogyakarta, Desember 2013

Yoseph Samuel

INTISARI

Implementasi Algoritma *K-Nearest Neighbor* dengan *Decision Rule* untuk Subtopik Berita

Klasifikasi dokumen sering digunakan oleh instansi untuk merapikan tiap dokumen yang mereka miliki agar dapat mudah ditemukan. Berita merupakan sebuah dokumen kecil yang jumlahnya banyak, sehingga perlu dilakukan klasifikasi untuk mempermudah pengaksesan data. Banyak berita yang sudah terklasifikasi berdasarkan topik berita. Karena satu topik berita dapat memuat banyak artikel,maka diperlukan juga klasifikasi berdasarkan subtopik beritanya.

Proses pengklasifikasian subtopik berita melalui beberapa tahapan, pertama berita yang ingin diproses harus melewati tahap *preprocess*ing yang meliputi *tokenizing*, *stopword*, *stemming* dan *sorting.* selanjutnya dari hasil *preprocess*ing, diterapkan Algoritma K-Nearest Neighbor untuk menentukan pengklasifikasian artikel. Dengan bantuan pembobotan *TF-IDF* dan pencarian kesamaan menggunakan *cosine* dan *euclidean distance* sebagai *Decision Rule*, maka akan didapati hasil klasifikasi subtopik.

Hasil dari penelitian ini, klasifikasi subtopik berita sudah baik, dengan persentase paling tinggi yaitu 89,36%, diketahui bahwa hasil tersebut merupakan hasil dengan k = 3 dan menggunakan Decision Rule. Penggunaan Decision Rule sebenarnya hanya menambahkan 1,07% dari hasil sebelumnya yaitu 88,29%.

Kata kunci : klasifikasi, *K-Nearest Neighbor, Decision Rule, Stemming, Tokenizing, TF-IDF, cosine, Euclidean distance*

DAFTAR ISI

HALAMAN JUDUL ……
PERNYATAAN KEASLIAN SKRIPSI ……...… iii
HALAMAN PERSETUJUAN ……..... iv
HALAMAN PENGESAHAN ……..…… v
UCAPAN TERIMA KASIH ……..….vi
KATA PENGANTAR ……..… vii
INTISARI ……...……... viii
DAFTAR ISI ……..…….....…..ix
DAFTAR TABEL ……..xi
DAFTAR GAMBAR ……..……. xii

Bab 1 PENDAHULUAN 1

1.1 Latar Belakang Masalah 1

1.2 Perumusan Masalah 2

1.3 Batasan Masalah 2

1.4 Tujuan Penelitian 3

1.5 Metode Penelitian 3

1.6 Sistematika Penulisan 4

Bab 2 TINJAUAN PUSTAKA 5

2.1 Tinjauan Pustaka 5

2.2 Landasan Teori 6

2.2.1 *Preprocess*ing 6

2.2.2 Pembobotan 8

2.2.3 Klasifikasi 9

BAB 3 ANALISIS DAN PERANCANGAN SISTEM 15

3.1. Alat Penelitian 15

3.1.1. Perangkat Keras 15

3.1.2. Perangkat Lunak 15

3.2. Rancangan Sistem 16

3.2.1. Diagram Alir (*flowchart*) 16

3.2.2. Perancangan Basis Data (skema diagram) 21

3.2.3. Algoritma Program 22

3.2.4. Perancangan Antarmuka Sistem 23

3.2.5. Perancangan Pengujian Sistem 26

Bab 4 IMPLEMENTASI DAN ANALISIS SISTEM 28

4.1. Implementasi Sistem 28

4.1.1 Antarmuka Program 28

4.1.2 Implementasi Algoritma 34

4.1.3 Pengumpulan Berita 37

4.2. Analisis Sistem 38

4.2.1 Evaluasi Keakuratan Sistem 39

4.2.2 Evaluasi Pengaruh Hasil Stemming 43

BAB 5 KESIMPULAN DAN SARAN 45

5.1. Kesimpulan 45

5.2. Saran 46

DAFTAR TABEL

Tabel 4.1 Persentase Keakuratan : K-NN untuk berita BBC.....…… 42
Tabel 4.2 Persentase Keakuratan : K-NN untuk berita CNN…....…….... 42
Tabel 4.3 Persentase Keakuratan : K-NN untuk berita FOX .....……..…… 43
Tabel 4.4 Data Berita_Training : 5 term yang paling banyak muncul…..... 43
Tabel 4.5 Berita_Baru : 5 term ID_berita = 10…….. 43
Tabel 4.6 Persentase Keakuratan : Algoritma K-NN with Decision Rule BBC..... 44
Tabel 4.7 Persentase Keakuratan : Algoritma K-NN with Decision Rule CNN.....44
Tabel 4.8 Persentase Keakuratan : Algoritma K-NN with Decision Rule FOX..... 45
Tabel 4.9 Hasil Euclidean : id_berita 3 dan sumber foxnews(3)..... 46
Tabel 4.10 Hasil cosine : id_berita 3 dan sumber foxnews(3)..... 47
Tabel 4.11 Persentase Keakuratan : Algoritma K-NNwDR dan K-NN..... 48

DAFTAR GAMBAR

Gambar 2.1 K-Nearest Neighbor dengan 3 neighbor..... 13
Gambar 3.1 Diagram Alir Utama..... 19
Gambar 3.2 Diagram Alir *Preprocess*ing.....20
Gambar 3.3 Diagram Alir Pembobotan TF-IDF..... 21
Gambar 3.4 Diagram Alir Algoritma K-Nearest Neighbor..... 22
Gambar 3.5 Diagram Alir Decision Rule menggunakan Euclidean Distance..... 23
Gambar 3.6 Diagram Basis Data..... 24
Gambar 3.7 Rancangan Halaman Utama User..... 26
Gambar 3.8 Rancangan Halaman Add Berita..... 27
Gambar 3.9 Rancangan Halaman Cek Berita..... 28
Gambar 3.10 Rancangan Halaman Hasil..... 29
Gambar 4.1 Halaman Verifikasi..... 31
Gambar 4.2 Halaman User..... 32
Gambar 4.3 Halaman Admin Login.....33
Gambar 4.4 Halaman Admin Add Berita..... 34
Gambar 4.5 Halaman Admin Cek Berita..... 35
Gambar 4.6 Halaman Admin Hasil Berita..... 36
Gambar 4.7 Perhitungan Sum..... 38
Gambar 4.8 Perhitungan Length dan Cosine..... 38
Gambar 4.9 Perhitungan Euclidean Distance.....39
Gambar 4.10 Perhitungan Decision Rule.....40

INTISARI

Implementasi Algoritma *K-Nearest Neighbor* dengan *Decision Rule* untuk Subtopik Berita

Klasifikasi dokumen sering digunakan oleh instansi untuk merapikan tiap dokumen yang mereka miliki agar dapat mudah ditemukan. Berita merupakan sebuah dokumen kecil yang jumlahnya banyak, sehingga perlu dilakukan klasifikasi untuk mempermudah pengaksesan data. Banyak berita yang sudah terklasifikasi berdasarkan topik berita. Karena satu topik berita dapat memuat banyak artikel,maka diperlukan juga klasifikasi berdasarkan subtopik beritanya.

Proses pengklasifikasian subtopik berita melalui beberapa tahapan, pertama berita yang ingin diproses harus melewati tahap *preprocess*ing yang meliputi *tokenizing*, *stopword*, *stemming* dan *sorting.* selanjutnya dari hasil *preprocess*ing, diterapkan Algoritma K-Nearest Neighbor untuk menentukan pengklasifikasian artikel. Dengan bantuan pembobotan *TF-IDF* dan pencarian kesamaan menggunakan *cosine* dan *euclidean distance* sebagai *Decision Rule*, maka akan didapati hasil klasifikasi subtopik.

Hasil dari penelitian ini, klasifikasi subtopik berita sudah baik, dengan persentase paling tinggi yaitu 89,36%, diketahui bahwa hasil tersebut merupakan hasil dengan k = 3 dan menggunakan Decision Rule. Penggunaan Decision Rule sebenarnya hanya menambahkan 1,07% dari hasil sebelumnya yaitu 88,29%.

Kata kunci : klasifikasi, *K-Nearest Neighbor, Decision Rule, Stemming, Tokenizing, TF-IDF, cosine, Euclidean distance*

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Berita merupakan sebuah alat yang digunakan oleh manusia dari zaman ke zaman. Berita memberikan informasi bagi manusia. Informasi yang diberikan tentulah aktual dan terpercaya. Setiap manusia pasti selalu mencari informasi yang terpercaya dan aktual setiap harinya. Dimana dari sifat manusia sendiri selalu ingin mengetahui dan sifat dari berita selalu memberikan informasi. Manusia sangatlah lekat dengan berita, maka informasi yang diberikan pastinya terpercaya dan aktual.

Banyak surat kabar *online* telah membuat desain webnya semakin mudah dilihat oleh *user*. Pengklasifikasian berita menjadi topik-topik berita membuat pencarian berita dipermudah. Beberapa surat kabar *online* memberikan sebuah fitur baru yaitu pengklasifikasian berdasarkan subtopik. Bagi *user*, klasifikasi tersebut memberikan kemudahan agar saat *user* mencari berita, berita dapat dicari melalui kategori yang lebih detail. Sedikit dari surat kabar *online* menggunakan fitur tersebut.

Penggunaan *K-Nearest Neighbor* sangat umum digunakan untuk pengkategorisasian teks. Hal tersebut diketahui karena algoritmanya yang mudah dan efisien untuk klasifikasi teks. Bukan hanya dari mudah dan efisien, sifat dari algoritma *K-Nearest Neighbor* sendiri merupakan *self-learning* dimana algoritma ini dapat mempelajari struktur data yang ada dan menkategorikan dirinya. Biasanya, *K-Nearest Neighbor* selalu menggunakan *majority vote* sebagai landasan penentuan dimana sebuah dokumen diklasifikasi. Permasalahannya adalah jika ada sebuah kategori dimana kategori tersebut sudah mempunyai banyak dokumen, maka

kemungkinan besar yang terjadi jika ada dokumen baru dan mendekati kategori tersebut akan ikut tergeser kedalam kategori karena penggunaan *majority vote*. Disini penulis akan mengganti penggunaan *majority vote* menjadi sebuah *Decision Rule* yang lebih baik agar penggunaan algoritma *K-Nearest Neighbor* dapat dimaksimalkan.

1.2 Perumusan Masalah

Berdasarkan latar belakang masalah diatas, penulis akan merancang dan membangun sebuah sistem yang akan melakukan proses klasifikasi untuk subtopik berita. Masalah yang akan diteliti lebih lanjut adalah :

1. Dengan menggunakan algoritma *K-Nearest Neighbor*, apakah dapat mengklasifikasi berita menjadi subtopik dari topik berita olahraga yang sudah ditetapkan?
2. Dengan menggunakan Decision Rule, apakah berita yang terklasifikasi lebih akurat dibandingkan dengan algoritma *K-Nearest Neighbor* saja?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini, yaitu :.

1. Pengambilan berita dibatasi menjadi 3 alamat *website* yaitu : *bbc.com*, *cnn.com*, *foxnews.com*
2. Berita yang digunakan dikategorikan berdasarkan topik olahraga dan terbagi menjadi 7 subtopik (*Soccer, Formula 1, Basketball, Motorsport, Baseball, Tennis, NFL*)
3. Berita yang digunakan adalah berita dalam bahasa Inggris.
4. Penggunaan kata hanya menggunakan huruf. Angka akan dihilangkan.
5. Berita yang ditampilkan adalah berita yang diambil sebagai data set. Data berita akan diambil 280 untuk data training yang diambil dari sumber

yang sudah ada dan 45 berita dari ketiga sumber yang telah ada dan 49 berita dari sumber lain (*cbssport.com, nytimes.com, theguardian.com, nbcsport.com*).

6. Diasumsikan sistem pengklasifikasi yang dilakukan oleh surat kabar online adalah benar.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengklasifikasi berita menggunakan algoritma *K-Nearest Neighbor with Decision Rule*.

1.5 Metode Penelitian

Metodologi yang akan dipakai dalam penelitian ini adalah :

1. Studi Pustaka

   Studi Pustaka dilakukan dengan mempelajari teori – teori melalui buku, artikel, jurnal dan bahan lain yang mendukung yang berhubungan dengan data mining, algoritma *Improved K-Nearest Neighbor*, dan metode – metode pendukung lainnya yang dibutuhkan.

2. Perancangan Sistem

   Pada tahap ini sistem yang akan dirancang didasarkan pada proses yang berlaku. Proses akan berlaku diawal dimana algoritma Improved *K-Nearest Neighbor* digunakan untuk mengklasifikasi tiap berita yang muncul dan akan terbagi menjadi 4 pilihan tab. User tidak akan diberitahu tentang proses tersebut dan langsung dapat menggunakan fitur pembacaan berita.

3. Pembangunan Sistem

Pada tahap ini program akan dibuat disesuaikan dengan rancangan sistem.

4. Implementasi dan Testing

   Pengujian terhadap program dengan melihat apakah semua berita yang ada dapat terklasifikasi dengan benar di subtopik - subtopiknya masing - masing

5. Analisis Hasil Percobaan dan Evaluasi

   Pada tahap ini kesimpulan dapat ditariik setelah melakukan uji coba pada program.

1.6 Sistematika Penulisan

Penulisan laporan skripsi ini dibagi menjadi lima (5) bab, yaitu :

Bab 1 PENDAHULUAN yang berisi latar belakang masalah, perumusan masalah, batasan masalah, hipotesis, tujuan penelitian, metodologi, dan sistematika penulisan Skripsi.

Bab 2 TINJAUAN PUSTAKA yang berisi gagasan-gagasan yang muncul dengan memberikan landasan teori yang akurat dari berbagai sumber dan konsep-konsep yang dibutuhkan dalam pengertian *K-Nearest Neighbor* dengan *Decision Rule*

Bab 3 ANALISIS DAN PERANCANGAN SISTEM yang berisi perancangan sistem yang akan memberikan gambaran sistem yang akan dibuat serta prosedur–prosedur yang digunakan dalam sistem.

Bab 4 IMPLEMENTASI DAN ANALISIS SISTEM yang berisi implementasi dari hasil perancangan sistem dan pengujian terhadap sistem yang telah dibuat.

Bab 5 KESIMPULAN DAN SARAN yang berisi kesimpulan atas sistem yang telah dibuat serta saran-saran dalam pengembangan dari Skripsi ini agar dapat dikembangkan kembali.

BAB 5

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan maka dapat disimpulkan :

1. Penggunaan K-Nearest Neighbor sebagai klasifikasi menunjukan persentasi yang baik, dengan nilai k = 3, menunjukan hasil persentase 88,29%. Dari k yang sama, digunakan Decision Rule yang ada dan persentase hasil akhir dari keakuratan K-Nearest Neighbor dengan Decision Rule adalah 89,36%.
2. Dari kesimpulan tersebut maka dapat disimpulkan menggunakan k = 3 merupakan k yang paling tinggi keakuratannya dalam K-Nearest Neighbor maupun K-Nearest Neighbor with Decision Rule.
3. Penggunaan Decision Rule hanya akan menambah keakuratan yang sedikit dan kurang mampu memaksimalkan performa K-Nearest Neighbor sendiri. Dari Algoritma K-Nearest Neighbor sendiri sudah memuaskan karena hasilnya yang tinggi.
4. TF.IDF selalu akan digunakan sebagai pembobotan K-NN dengan diingat bahwa ada word count bonus dimana hasil IDF akan ditambahkan dengan 1. Selain itu, hasil Euclidean Distance sebagai Decision Rule merupakan hasil yang signifikan dan jelas, memungkinkan penulis untuk memberikan hasil yang lebih baik.

5.2. Saran

Sistem yang digunakan merupakan klasifikasi terhadap subtopik berita, maka dari itu dalam pengembangan sistem kedepan,

1. Hasil *preprocess* yang lebih baik, dimisalkan penambahan stopword dari link dan buku yang sudah ada, stemming yang mempunyai library sendiri, hasil stemming yang diharapkan tidak membuat keambiguan dalam sebuah proses.
2. Tidak menutup kemungkinan untuk menambahkan atau mengganti sistem dengan Decision Rule baru dan Algoritma baru.

DAFTAR PUSTAKA

Francis, A.L.,FCAAS, MAAA. (2006). *Taming text : An introduction to Text Mining. Casuality Acutuarial Society Forum*

Grosman, A.D., & Frieder,). (2004). *Information Retrieval : Algorithms and Heuristics.* Netherland : Springer, Inc

Han, E., Karypis, & Kumar. (2001) *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification.* Journal Department of Computer Science and Engineering Army HPC Research Center. University of Minnesota.

Herwansyah, A. (2009). Aplikasi Pengkategorian Dokumen dan Pengukuran Tingkan Similaritas Dokumen Menggunakan kata Kunci pada Dokumen penulisan Ilmiah. Jurnal Sistem Informasi. Universitas Gunadarma.

Miah, M. (2009). *Improved k-nn Algorithm for Text Classification.* Journal Department of Science and Engineering. University of Texas.

Robertson, S. (2004). *Understanding Inverse Document Frequency : On theorethical argument for IDF* Journal of Documentation 60 no. 5, pp 503-520. Cambridge.

Weiss, M.S., Indurkhya, Zhang & Damerau. (2005). *Text Mining : Predictive Methods for Analyzing Unstructured Information.* New York: Springer, Inc.