KLASIFIKASI JENIS QUOTE DENGAN K-NEAREST NEIGHBOR

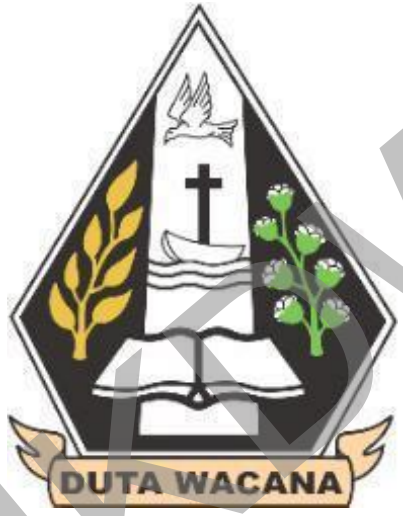
Skripsi

oleh
KEVIN ADRIAN WAHYU NUGROHO
22094698

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI UNIVERSITAS KRISTEN DUTA WACANA 2017

KLASIFIKASI JENIS QUOTE DENGAN K-NEAREST NEIGHBOR

Skripsi



Diajukan kepada Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

KEVIN ADRIAN WAHYU NUGROHO
22094698

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS
TEKNOLOGI INFORMASI UNIVERSITAS KRISTEN DUTA WACANA
2017

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

KLASIFIKASI JENIS QUOTE DENGAN K-NEAREST NEIGHBOR

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, 1 Agustus 2017

KEVIN ADRIAN WAHYU NUGROHO
22094698

HALAMAN PERSETUJUAN

Judul Skripsi	:	Klasifikasi Jenis Quote Dengan K-Nearest Neighbor
Nama	:	Kevin Adrian Wahyu Nugroho
N I M	:	22094698
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Sisipan
Tahun Akademik	:	2016/2017

Telah diperiksa dan disetujui
di Yogyakarta,
Pada tanggal 6 Juli 2017

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat C., S.Kom., M.Cs	Budi Susanto, S.Kom., M.T

HALAMAN PENGESAHAN

KLASIFIKASI JENIS QUOTE DENGAN K-NEAREST NEIGHBOR

Oleh: KEVIN ADRIAN WAHYU NUGROHO / 22094698

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 1 Agustus 2017

Yogyakarta, 1 Agustus 2017
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C., S.Kom.,M.Cs.
2. Budi Susanto, SKom.,M.T.
3. Restyandito, S.Kom.,MSIS, Ph.D
4. Yuan Lukito, S.Kom., M.Cs.

Dekan	Ketua Program Studi
(Budi Susanto, S.Kom., M.T.)	(Gloria Virginia, Ph.D.)

UCAPAN TERIMA KASIH

Puji dan syukur panjatkan kepada Tuhan Yesus Kristus atas anugerah dan penyertaannya sehingga dapat terselesaikannya Tugas Akhir dengan judul Klasifikasi Jenis Quote Dengan K-Nearest Neighbor

Penulisan laporan merupakan tugas kelengkapan dan salah satu syarat yang diperlukan untuk memperoleh gelar Sarjana Komputer (S1), dan juga melatih mahasiswa untuk membuat suatu karya ilmiah yang dapat dipertanggung jawabkan secara ilmiah.

Dalam menyelesaikan pembuatan program dan laporan Tugas Akhir ini, penulis mendapat banyak bimbingan, saran, masukan, dan semangat moral dari berbagai pihak baik secara langsung maupun tidak langsung. Untuk itu dengan segala kerendahan hati, pada kesempatan ini penulis juga menyampaikan ucapan terima kasih kepada :

1. Tuhan Yesus Kristus atas uluran tangannya yang tidak pernah henti-hentinya serta pengharapan yang selalu disediakan-Nya.
2. Bapak Antonius Rachmat C., S.Kom.,M.Cs. selaku dosen pembimbing I yang telah banyak membantu, memberikan bimbingannya dengan sabar dan baik, juga memberikan petunjuk serta memberikan semangat kepada penulis dalam menyelesaikan skripsi ini.
3. Bapak Budi Susanto, S.Kom.,M.T. selaku dosen pembimbing II atas bimbingan, arahan, masukan, serta semangat kepada penulis selama pengerjaan skripsi ini dari awal hingga akhir.
4. Keluarga, papa, mama, dan adik yang selalu memberikan dukungan, semangat dan doa bagi penulis selama pengerjaan program dan laporan supaya dapat terselesaikan dengan baik dan tepat waktu. Terima kasih atas penantian yang tidak sebentar.
5. Sahabat-sahabat yang selalu memberikan semangat dan doa dalam pengerjaan Tugas Akhir ini

6. Teman-teman seperjuangan TI angkatan 2009 yang tidak bisa disebutkan satu persatu yang selalu bersama-sama berjuang untuk menyelesaikan Tugas Akhir.
7. Pihak lain yang tidak dapat penulis sebutkan satu per satu hingga terselesaikan Tugas Akhir ini dengan baik.

Penulis menyadari bahwa penulisan Tugas Akhir ini masih jauh dari sempurna, oleh karena itu penulis mengaharapkan kritik dan saran yang membangun dari pembaca sekalian. Sehingga suatu saat penulis dapat memberikan karya yang lebih baik.

Akhir kata penulis ingin meminta maaf, apabila terdapat kesalahan baik dalam penyusunan laporan maupun yang pernah penulis lakukan sewaktu membuat Tugas Akhir.

Yogyakarta, 14 Juli 2017

Penulis

INTISARI

KLASIFIKASI JENIS QUOTE DENGAN K-NEAREST NEIGHBOR

Quote berasal dari Bahasa Inggris yang artinya mengutip. Berdasarkan arti tersebut, *quote* dapat diartikan sebagai pernyataan orang-orang yang terkenal. Banyak kata-kata yang pernah diucapkan oleh seseorang menjadi inspirasi dan motivasi bagi pembacanya. Selain dari ucapan langsung tokoh terkenal, quote juga seringkali diambil dari buku dan film. Salah satu metode yang dapat digunakan untuk pengklasifikasian adalah K-Nearest Neighbor. Metode K-Nearest Neighbor digunakan untuk pengkategorian teks. Sifat dari K-Nearest Neighbor sendiri yaitu algoritma ini dapat mempelajari struktur data yang ada dan mengkategorikan dirinya. Penggunaan tools Rapid Miner digunakan untuk analisa keakuratan hasil klasifikasi quote yang akan dilakukan.

Rapid Miner memiliki operator yang dapat digunakan sesuai kebutuhan. Tersedia berbagai macam operator yang dapat dilakukan untuk klasifikasi. Operator yang digunakan untuk klasfikasi mulai dari tahap pembacaan data, preprocessing, pembobotan TF-IDF, k-Nearest Neighbor, dan pengujian. Seluruh proses klasifikasi quote dilakukan menggunakan Rapid Miner dan dari hasil tersebut akan dilakukan analisis.

K-nearest Neighbor dapat digunakan untuk klasifikasi quote. Persentasi keakuratan tertinggi dengan menggunakan nilai k = 5 yaitu 72.5%. Metode Prune untuk Feature Selection dengan persentual menghasilkan persentasi keakuratan yang lebih baik yaitu 72.5% dibandingkan by ranking yang hanya menghasilkan 25%. Metode Stemming menggunakan Snowball dan Porter menghasilkan presentasi keakuratan sama dengan Lovins yaitu 72.5%.

Kata Kunci : klasifikasi, quote, *K- Nearest Neighbor*, Rapid Miner.

DAFTAR ISI

HALAMAN JUDUL .....
PERNYATAAN KEASLIAN SKRIPSI iii
HALAMAN PERSETUJUAN iv
HALAMAN PENGESAHAN v
UCAPAN TERIMAKASIH vi
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 2
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 Text Mining 6
2.2.1.1 Tokenisasi 7
2.2.1.2 Stopwords 7
2.2.1.3 Stemming 7
2.2.1.4 Pembobotan TF-IDF 7
2.2.2 Algoritma K-Nearest Neighbor 8
2.2.3 Confusion Matrix 10
2.2.4 RapidMiner 11

BAB 3 ANALISIS DAN PERANCANGAN SISTEM 1 3
3.1 Spesifikasi Sistem 13
3.1.1 Perangkat Keras 13
3.1.2 Perangkat Lunak 13
3.2 Rancangan Sistem 14
3.2.1 Diagram Alir (*Flowchart*) 14
3.2.2 Sumber Data 18
3.2.3 Rancangan Desain Proses Klasifikasi Sistem 19
3.2.4 Perancangan Pengujian Sistem 20
BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM 22
4.1 RapidMiner 22
4.2 Data Quote 23
4.3 Pembacaan Sumber Data 26
4.4 Implementasi Algoritma 30
4.5 Analisis Sistem 30
BAB 5 KESIMPULAN DAN SARAN 41
5.1 Kesimpulan 41
5.2 Saran 41
DAFTAR PUSTAKA 42
LAMPIRAN ..... A -1

DAFTAR TABEL

Tabel 2.1 Confusion Matrix……..... 10
Tabel 3.1 Tabel Sumber Data Quote……..... 19
Tabel 3.2 K-Fold Cross Validation……..……..……..... 21
Tabel 4.1 Hasil Pengujian dengan nilai k=1 hingga k=10……..... 36

DAFTAR GAMBAR

Gambar 2.1 Proses Text Mining 6
Gambar 2.2 K-Nearest Neighbor dengan 3 neighbor....., ... 9
Gambar 2.3 Tampilan RapidMiner……..…..... 12
Gambar 3.1 Diagram Alir Utama……..…..... 14
Gambar 3.2 Diagram Alir Preprocessing…….....15
Gambar 3.3 Diagram Alir Pembobotan TF-IDF…… 16
Gambar 3.4 Diagram Alir Algoritma K-Nearest Neighbor…….. 17
Gambar 3.5 Diagram Alir K-Fold Cross Validation…….....18
Gambar 3.6 Preprocessing pada RapidMiner……19
Gambar 3.7 Proses Klasifikasi dan Pengujian pada RapidMiner… 20
Gambar 4.1 Data Quote pada file Excel……..... 23
Gambar 4.2 Sumber Data Quote…….....24
Gambar 4.3 Jumlah Kemunculan Author…… 25
Gambar 4.4 Operator Read Excel…….....26
Gambar 4.5 Operator Process Documents from Data…… ..27
Gambar 4.6 Preprocessing…….....28
Gambar 4.7 Operator k-NN……..... 29
Gambar 4.8 Bagian Pengujian…….....29
Gambar 4.9 Hasil Preprocessing... ..…….....…31
Gambar 4.10 Hasil Pembobotan TF-IDF…….....31
Gambar 4.11 Hasil Pengujian dengan k=1……..……..... …….32
Gambar 4.12 Hasil Pengujian dengan k=2……..…….....32
Gambar 4.13 Hasil Pengujian dengan k=3……..…….....33
Gambar 4.14 Hasil Pengujian dengan k=4……..…….....33
Gambar 4.15 Hasil Pengujian dengan k=5……..…….....34
Gambar 4.16 Hasil Pengujian dengan k=6……..…….....34
Gambar 4.17 Hasil Pengujian dengan k=7……..…….....34
Gambar 4.18 Hasil Pengujian dengan k=8……..…….....35

Gambar 4.19 Hasil Pengujian dengan k=9……..……..... 35
Gambar 4.20 Hasil Pengujian dengan k=10……..……..... 36
Gambar 4.21 Grafik Persentase Hasil Pengujian…..……..... 37
Gam bar 4 . 22 P roses Stem ming m enggunakan al goritm a P ort er
dan Snowball....……..……..... 37
Gambar 4.23 Hasil Pengujian setelah Stemming menggunakan Snowball
dan Porter dengan nilai k=5...…… ……..... 38
Gambar 4.24 Metode Prune by ranking……..……..... 38
Gambar 4.25 Hasil Metode Prune by ranking……..... 39
Gambar 4.26 Hasil Pengujian Prune by ranking dengan k=5……..... 39

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Quote berasal dari Bahasa Inggris yang artinya mengutip. Berdasarkan arti tersebut, *quote* dapat diartikan sebagai pernyataan orang-orang yang terkenal. Banyak kata-kata yang pernah diucapkan oleh seseorang menjadi inspirasi dan motivasi bagi pembacanya. Selain dari ucapan langsung tokoh terkenal, quote juga seringkali diambil dari buku dan film. Beberapa website seperti www.brainyquote.com, www.great-quotes.com, www.quotegarden.com menyediakan *quote* dalam berbagai kategori seperti *Love, Inspirational, Friendship,* dan *Happiness.* Pengklasifikasian *quote* menjadi berbagai kategori membuat pencarian *quote* lebih mudah.

Salah satu metode yang dapat digunakan untuk pengklasifikasian adalah K-Nearest Neighbor. Metode K-Nearest Neighbor digunakan untuk pengkategorian teks. Sifat dari K-Nearest Neighbor sendiri yaitu algoritma ini dapat mempelajari struktur data yang ada dan mengkategorikan dirinya. Algoritma ini melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Metode ini sangat sederhana, mudah direpresentasikan, memiliki ketangguhan terhadap *training data* yang memiliki banyak *noise*, dan cukup efektif untuk proses pengelompokan (Tan *et al*., 2005).

Dalam penelitian ini penulis akan menggunakan tools RapidMiner untuk mengklasifikasi *quote* dengan metode *K-Nearest Neighbor*. Penggunaan tools RapidMiner digunakan untuk analisa keakuratan hasil klasifikasi quote yang akan dilakukan.

1.2 Perumusan Masalah

Berdasarkan latar belakang maka penelitian yang akan dilakukan adalah :

1. Apakah metode *K-Nearest Neighbor* dapat digunakan untuk melakukan klasifikasi quote dengan RapidMiner?
2. Bagaimana keakuratan hasil klasifikasi *quote* menggunakan algoritma *K-Nearest Neighbor* berdasarkan *Confusion Matrix*?

1.3 Batasan Masalah

Batasan – batasan masalah yang didefinisikan dalam penelitian ini adalah:

1. *Quote* yang digunakan adalah quote dalam bahasa Inggris.
2. Pengambilan *quote* melalui www.quoteland.com dan www.brainyquote.com.
3. Kategori yang digunakan adalah *Love, Inspirational, Friendship, dan Happiness,* masing-masing 50 quote yang akan digunakan sebagai data latih
4. Menggunakan algoritma Porter untuk stemming kata dasar yang sama
5. Stopwords list dalam bahasa Inggris.
6. Menggunakan tools RapidMiner (www.rapidminer.com)

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah menguji keakuratan metode K-Nearest Neighbor untuk mengklasifikasikan quote

1.5. Metode Penelitian

Metode yang akan digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi Pustaka

   Studi Pustaka dilakukan dengan mempelajari teori-teori melalui buku, artikel, jurnal dan bahan lain yang mendukung dan berhubungan dengan *text mining*, algoritma *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 *K-Nearest Neighbor* digunakan untuk mengklasifikasi *quote* Kemudian menyiapkan data yang dibutuhkan yaitu data latih dan data uji. Data latih untuk membangun pengklasifikasian quote dan data uji digunakan untuk menguji keakuratan sistem klasifikasi

3. Penggunaan Tools RapidMiner

   Pada tahap ini dilakukan proses klasifikasi quote dengan menggunakan tools RapidMiner.

4. Testing

   Pada tahap ini dilakukan pengujian dengan metode K-Nearest Neighbor untuk pengklasifikasian quote. Pengujian menggunakan tools RapidMiner dengan melihat apakah semua quote yang ada dapat terklasifikasi dengan benar.

5. Analisis hasil percobaan dan evaluasi

   Setelah dilakukan pengujian, tahap selanjutnya adalah menganalisis keefektifan algoritma *K-Nearest Neighbor* untuk pengklasifikasian quote dengan confussion matrix

1.6. Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini dapat dijabarkan sebagai berikut :

Bab I Pendahuluan berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, metode penelitan, serta sistematika penulisan.

Bab II Tinjauan Pustaka berisi landasan teori dan tinjauan pustaka. Landasan Teori memuat penjelasan tentang konsep dan prinsip utama yang digunakan untuk memecahkan masalah dari penelitian ini. Dan Tinjauan Pustaka berisi tentang berbagai hasil penelitian lainnya yang didapatkan dari sumber pustaka seperti jurnal ilmiah yang akan membantu peneliti dalam menguraikan berbagai teori pendukung penelitian.

Bab III Analisis dan Perancangan Sistem berisi perancangan proses, penjelasan tentang algoritma serta pengimplementasian metode *K-Nearest Neighbor*.

Bab IV Implementasi dan Analisis Sistem berisi hasil akhir dari pengujian klasifikasi quote menggunakan RapidMiner dan hasil evaluasi dari penelitian tersebut.

Bab V Kesimpulan dan Saran berisi kesimpulan dan saran dari keseluruhan penilitian yang telah dilakukan.

BAB 5

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan maka dapat disimpulkan :

1. K-Nearest Neighbor dapat digunakan untuk klasifikasi quote dengan menggunakan RapidMiner
2. Penggunaan K-Nearest Neighbor sebagai klasifikasi menunjukkan persentasi keakuratan tertinggi sebesar 72.5% dengan nilai k = 5.
3. Metode Prune untuk Feature Selection dengan persentual menghasilkan persentasi keakuratan yang lebih baik yaitu 72.5% dibandingkan by ranking yang hanya menghasilkan 25%
4. Metode Stemming menggunakan Snowball dan Porter menghasilkan presentasi keakuratan sama dengan Lovins yaitu 72.5%

5.2. Saran

Penelitian yang digunakan merupakan klasifikasi terhadap quote, maka dari itu dalam pengembangan kedepan dapat dilakukan,

1. Penambahan stopword dari link dan buku yang sudah ada. Stemming yang mempunyai library sendiri sehingga hasil stemming yang diharapkan tidak membuat keambiguan dalam sebuah proses.
2. Jumlah data yang digunakan dapat diperbanyak dan menggunakan bahasa Indonesia.
3. Menggunakan algoritma lain untuk klasifikasi yang tersedia sebagai operator di RapidMiner

DAFTAR PUSTAKA

Aprilla, D., Baskoro, D., Ambarwati, L., & Wicaksana, I. (2013). *Belajar Data Mining dengan RapidMiner*

Feldman, R.; Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

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

Kohavi, R. & Provost, F. (1998). Glossary of terms. *Machine Learning*, *30*(2-3), 271-274.

Kumar, S., & Karthika, R. (2014). A Survey On Text Mining Process And Techniques

Lovins, J. (1968). *Development of a stemming algorithm*. Cambridge: M.I.T. Information Processing Group, Electronic Systems Laboratory.

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

Porter, M.F. (1980), An algorithm for suffix stripping, *Program*, Vol. 14 No.3,

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

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston: Pearson Addison Wesley.

Tuba, P. A. L. A., & Camurcu, A. Y. (2014). Evaluation Of Data Mining Classification And Clustering Techniques for Diabetes