PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA

Skripsi

oleh
ANTON SUSILO
71140050

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2019

PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA

Skripsi

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

Disusun oleh

ANTON SUSILO
71140050

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2019

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi 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, 9 Januari 2019

ANTON SUSILO
71140050

HALAMAN PERSETUJUAN

Judul Skripsi	:	PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA
Nama Mahasiswa	:	ANTON SUSILO
N I M	:	71140050
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2018/2019

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 9 Januari 2019

Dosen Pembimbing I	Dosen Pembimbing II
Yuan Lukito, S.Kom., M.Cs.	Antonius Rachmat C., S.Kom.,M.Cs.

HALAMAN PENGESAHAN

PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA

Oleh: ANTON SUSILO / 71140050

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 14 Desember 2018

Yogyakarta, 9 Januari 2019
Mengesahkan,

Dewan Penguji:

1. Yuan Lukito, S.Kom., M.Cs. .....
2. Antonius Rachmat C., S.Kom.,M.Cs. .....
3. Willy Sudiarto Raharjo, S.Kom.,M.Cs. .....
4. Lucia Dwi Krisnawati, Dr. Phil. .....

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

ABSTRAKSI

PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA

Instagram menjadi salah satu sosial media yang sering dipakai untuk membagikan momen dari tiap penggunanya melalui foto. Banyak pula *public figure* yang menggunakan sosial media ini sebagai media berbagi mereka. Namun, popularitas dari artis ini membuat beberapa kalangan mengirimkan komentar *spam* yang membuat komentar dari *post* artis sendiri menjadi kotor dipenuhi *spam* tersebut, sehingga membingungkan saat ingin membaca komentar dari *post* mereka.

Penelitian mengenai klasifikasi teks untuk masalah ini dilakukan dengan pertama kali melakukan pelatihan kepada sistem terlebih dahulu dengan data latih yang jumlahnya diambil secara random. Setelah proses pelatihan, dilakukan pengujian berdasarkan data uji dan latih dengan beberapa parameter seperti nilai k dan presentase fitur yang akan digunakan untuk menguji dan membandingkan metode KNN maupun DWKNN dan menampilkan hasil klasifikasi dari tiap metode dalam bentuk yang mudah dipahami.

Hasil penelitian menunjukkan bahwa pengubahan nilai k ini tidak memiliki dampak yang terlalu berarti dalam klasifikasi dengan metode DWKNN dalam penelitian ini, berbanding terbalik dengan KNN yang nilainya cenderung menurun seiringin penambahan nilai k dilakukan dan seleksi fitur tidak perlu digunakan dikarenakan hasil *success rate* yang lebih baik untuk presentase 80% hingga 100%.

Kata Kunci: Klasifikasi Teks, K-Nearest Neighbor, Distance Weighted K-Nearest Neighbor

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAKSI

PENERAPAN DISTANCE WEIGHTED K-NEAREST NEIGHBOR UNTUK KLASIFIKASI KOMENTAR INSTAGRAM BERBAHASA INDONESIA

Instagram menjadi salah satu sosial media yang sering dipakai untuk membagikan momen dari tiap penggunanya melalui foto. Banyak pula *public figure* yang menggunakan sosial media ini sebagai media berbagi mereka. Namun, popularitas dari artis ini membuat beberapa kalangan mengirimkan komentar *spam* yang membuat komentar dari *post* artis sendiri menjadi kotor dipenuhi *spam* tersebut, sehingga membingungkan saat ingin membaca komentar dari *post* mereka.

Penelitian mengenai klasifikasi teks untuk masalah ini dilakukan dengan pertama kali melakukan pelatihan kepada sistem terlebih dahulu dengan data latih yang jumlahnya diambil secara random. Setelah proses pelatihan, dilakukan pengujian berdasarkan data uji dan latih dengan beberapa parameter seperti nilai k dan presentase fitur yang akan digunakan untuk menguji dan membandingkan metode KNN maupun DWKNN dan menampilkan hasil klasifikasi dari tiap metode dalam bentuk yang mudah dipahami.

Hasil penelitian menunjukkan bahwa pengubahan nilai k ini tidak memiliki dampak yang terlalu berarti dalam klasifikasi dengan metode DWKNN dalam penelitian ini, berbanding terbalik dengan KNN yang nilainya cenderung menurun seiringin penambahan nilai k dilakukan dan seleksi fitur tidak perlu digunakan dikarenakan hasil *success rate* yang lebih baik untuk presentase 80% hingga 100%.

Kata Kunci: Klasifikasi Teks, K-Nearest Neighbor, Distance Weighted K-Nearest Neighbor

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Internet saat ini sudah menjadi salah satu bagian dari hidup manusia di era modern. Beberapa kegiatan sehari-hari sudah bisa dilakukan dengan Internet, seperti misalnya berbincang dengan teman-teman melalui aplikasi *chatting*, berbelanja, dan juga kegunaan lainnya yang salah satunya adalah mengirim pesan melalui e-mail. Namun, dari sekian banyak email yang dikirimkan dari internet, masih ada pesan email yang tidak valid dan dikirimkan secara terus menerus dengan alasan tertentu yang dikenal dengan istilah spamming. Masih ada beberapa oknum yang sengaja melakukan tindakan ini di Internet untuk beberapa keperluan baik pribadi maupun kelompok.

Spamming sendiri tidak hanya dikirimkan melalui email secara langsung, namun juga bisa dikirimkan melalui cara lain selain melalui email sebagai sarana pengirimannya, seperti misalnya pada kolom komentar Sosial Media. Sosial media yang semakin banyak digunakan oleh setiap orang saat ini seperti Facebook, Snapchat, Instagram, dan lainnya sudah menjadi sarana bagi para pengirim spam untuk mengirimkan informasi yang ingin disebar sebagai *spam* itu sendiri.

Instagram menjadi salah satu sosial media yang populer dikarenakan sekarang sudah semakin banyak pengguna Instagram dengan jumlah penggunanya yang sudah mencapai 1 Miliar Pengguna dengan sebanyak 95 juta posting setiap harinya (Clarke, 2018). Pertumbuhan pengguna dan juga besarnya jumlah penggunaan dari sosial media Instagram ini akan mendorong sebagian orang untuk melakukan *spamming* di sosial media tersebut dengan harapan semakin banyaknya orang yang melihat pesan *spam* dan semakin efektif pula pengiriman *spam* tersebut dengan memanfaatkan banyak pengguna.

Fitur *filter spam* pada Instagram tentu diperlukan dengan fungsi agar pengguna yang memiliki banyak penonton atau pengikut (seperti artis, *public figure*, dkk) akan merasa aman dan lebih mudah untuk mencari apakah komentar

yang mengisi posting yang mereka punya lebih bersih dan efisien sehingga lebih mudah bagi pemilik *post* untuk berinteraksi dengan komentar yang sebenarnya. Fitur *filter spam* seperti ini juga tidak lepas dari proses yang dinamakan dengan klasifikasi teks, sehingga dapat dibuat sistem yang bisa memberi kelas kepada dokumen baru berdasarkan data yang sudah dilatihkan sebelumnya pada sistem.

1.2 Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah seberapa tingkat keakuratan *Distance Weighted* K-Nearest Neighbor jika dibandingkan dengan k-Nearest Neighbor pada umumnya dalam melakukan pengecekan komentar *spam* pada situs Instagram

1.3 Batasan Masalah

Pada penelitian ini, peneliti akan menggunakan *corpus* berbahasa Indonesia, dengan memilih komentar dari beberapa *post* yang dibuat oleh *public figure* Indonesia (Chrismanto, Raharjo, & Lukito, 2018) yang terdiri dari 3148 data latih yang terdiri dari 988 data bukan spam serta 2160 data spam dengan menggunakan 516 data uji yang terdiri dari 270 data bukan spam dan 246 data yang merupakan spam, dengan melalui proses *Stemming* dengan menggunakan *Library* milik Sastrawi dan menggunakan Bahasa Indonesia.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menerapkan algoritma *Distance Weighted K-Nearest Neighbor* pada sistem klasifikasi komentar Instagram, serta meneliti seberapa akurat algoritma *Distance Weighted K-Nearest Neighbor* jika dibandingkan dengan algoritma *K-Nearest Neighbor* yang sudah ada dengan melakukan penelitian dari beberapa parameter, seperti jumlah K itu sendiri untuk tiap algoritmanya dan juga membandingkan nilai k yang optimal melalui seleksi fitur.

1.5 Metodologi Penelitian

Metode yang digunakan dalam penelitian ini antara lain sebagai berikut:

- Studi Literatur

  Studi Pustaka dilakukan dengan mempelajari teori-teori yang berkaitan dengan Data Mining melalui beberapa media seperti buku, jurnal ilmiah, serta sumber lainnya yang mendukung

- Pengumpulan Data

  Pengumpulan data dilakukan dengan mengambil *dataset* dari penelitian yang sudah dilakukan sebelumnya (Chrismanto, Raharjo, & Lukito, 2018)

- Pembuatan Sistem

  Sistem yang dibuat merupakan prototype dan dibuat dengan bahasa pemrograman *website* dengan menggunakan bahasa Pemrograman PHP.

- Metode Pengujian

  Pengujian terhadap sistem dilakukan dengan menguji data baru yang tidak merupakan salah satu bagian dari *corpus* yang sudah diteliti dan lalu membandingkan hasil pengecekan *spam* dari kedua algoritma yang digunakan dalam penelitian ini, yaitu K-Nearest Neighbor biasa dengan *Distance Weighted K-Nearest Neighbor*.

1.6 Sistematika Penulisan

Pada Bab 1 berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian dan sistematika penulisan. Sub-bab pertama dari Bab 1 membahas mengenai latar belakang masalah komentar *spam* yang ada pada situs Instagram, kemudian pada Sub-bab kedua akan dirumuskan poin-poin masalah yang akan diselesaikan. Batasan-batasan sistem yang dibuat akan dijelaskan pada Sub-bab ketiga dan dilanjutkan dengan tujuan serta metode yang akan dilakukan dalam penelitian.

Pada Bab 2 berisi tinjauan pustaka dan landasan dari perancangan sistem. Bab ini juga menjelaskan tentang hal-hal yang mendukung pembuatan sistem

analisa cek komentar *spam* dari situs Instagram, termasuk didalamnya terdapat penjelasan tentang algoritma dan juga langkah-langkah lain yang dibutuhkan.

Pada Bab 3 berisi perancangan sistem, dimulai dari daftar kebutuhan sistem yang akan dibuat, struktur atau cara kerja sistem yang dijelaskan dengan *flow diagram*, kebutuhan sistem akan perangkat *hardware* maupun *software*, dan perancangan pengujian sistem.

Pada Bab 4 berisi hasil penelitian dari penerapan sistem Distance Weighted K-Nearest Neighbor, dimana akan menunjukkan persentase keakuratan dalam menentukan hasil dari *test case* yang akan digunakan dengan menggunakan tabel. Selain itu, juga akan dilakukan perbandingan dengan metode KNN dengan DWKNN dan lalu menganalisa hasil dari keduanya berdasarkan dari beberapa point penelitian, seperti nilai K serta seleksi fitur untuk tiap hasil dan algoritmanya.

Pada Bab 5 berisi kesimpulan dari apa yang sudah dibahas pada bab-bab sebelumnya dan sekaligus menjawab apa yang menjadi permasalahan terutama pada Bab 1. Jika penulis memiliki ide untuk penelitian selanjutnya, penulis dapat mencantumkan sub-bab tentang saran yang berisi rujukan penelitian lanjutan untuk mengembangkan sistem yang sudah dibuat.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis sistem yang dilakukan, maka dapat disimpulkan bahwa:

1. Sistem memiliki tingkat *success rate* yang lebih tinggi dengan menggunakan rumus DWKNN daripada KNN dalam melakukan klasifikasi komentar spam yang ada. Proses klasifikasi dibuat berdasarkan data latih yang sudah dimasukkan ke sistem sebelumnya dengan menggunakan data dari penelitian Chrismanto, Raharjo, dan Lukito pada tahun 2018 yang sudah dilabeli. DWKNN memiliki *success rate* dengan rata-rata sebesar 91.24% dan KNN memiliki *success rate* dengan rata-rata 84.88% dengan nilai k optimal di k = 1 dengan rata-rata yang sama, yaitu sebesar 91.86%.
2. Pada bagian Precision dan Recall, DWKNN dan KNN memiliki nilai Precision yang cenderung masih tinggi, namun Recall KNN pada penelitian ini memiliki presentase yang cenderung lebih rendah dan menandakan hasil klasifikasi yang kurang baik dikarenakan data latih yang tidak seimbang, walaupun DWKNN tidak terlalu banyak mendapat pengaruh dengan Recall yang cenderung stabil. Recall terendah KNN berada pada k = 9 dengan presentase 66.26%, dengan nilai Recall terendah DWKNN berada pada k = 3 dengan presentase 85.77%. Precision terendah KNN berada pada k = 5 dengan presentase 82.53% dan Precision terendah DWKNN berada pada k = 7 dengan 91.6%.
3. Berdasarkan nilai k optimal yang ditemukan, seleksi fitur pada sistem ini dengan menggunakan nilai k optimal memiliki hasil yang bagus pada nilai presentase fitur saat mencapai 80% dan 100% dengan tingkat *success rate* sebesar 91.86%

5.2 Saran

Sistem klasifikasi komentar Instagram ini masih bisa dikembangkan lebih baik lagi. Penulis memberikan saran sebagai berikut:

1. Perlu adanya kamus khusus untuk menyimpan kata tidak baku dan harus dikembangkan terus menyesuaikan komentar pengguna yang ada
2. Perlu adanya kamus lengkap untuk merubah emoji ataupun simbol dari bahasa lain seperti bahasa latin dan translasi bahasa asing ke Indonesia untuk bisa diproses menjadi kata baku dengan bahasa Indonesia lebih baik lagi, mengingat pengguna Instagram yang gemar menggunakan campuran Bahasa Indonesia ataupun Bahasa Asing.

DAFTAR PUSTAKA

Chrismanto, A. R., Raharjo, W. S., & Lukito, Y. (2018). Design and Development of REST-based Instagram Spam Detector for Indonesian Language.

Clarke, T. (2018, October 5). *24+ Instagram Statistics That Matter to Marketers in 2019*. Retrieved from 24+ Instagram Statistics That Matter to Marketers in 2019: https://blog.hootsuite.com/instagram-statistics/

Dudani, S. A. (1976, April). The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-6*(4), 325-327.

Gou, J., Du, L., Zhang, Y., & Xiong, T. (2012). A New Distance-weighted k-nearest Neighbor Classifier. *Journal of Information & Computational Science 9(6)*, 1429-1436.

Salton, G., & Buckley, C. (1988, January). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*.

Sun, S., & Huang, R. (2010). An Adaptive k-Nearest Neighbor Algorithm. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 1*, 91-94.

Suyanto. (2017). *Data Mining Untuk Klasifikasi dan Klasterisasi Data.* Informatika Bandung.

Wu, J., Cai, Z., & Gao, Z. (2010). Dynamic K-Nearest-Neighbor with Distance and Attribute Weighted for Classification. *2010 International Conference on Electronics and Information Engineering, 1*, 356.

Zaki, M. J., Jr., W. M., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms.* Cambridge University Press.