Edu Komputika 4 (1) (2017)

Edu Komputika Journal

http://journal.unnes.ac.id/sju/index.php/edukom

Pembuatan Aplikasi Klasifikasi Otomatis Laporan Keluhan Warga di Polrestabes Semarang

Dawam Muhammad Akbar✉ **Feddy Setio Pribadi**

Universitas Negeri Semarang

Info Artikel

Sejarah Artikel:
Diterima Juni 2016
Disetujui Juni 2016
Dipublikasikan Agustus 2017

Keywords:
classification, nearest neighbor, cosine similarity

Abstrak

Penelitian ini bertujuan untuk mempermudah dalam pengklasifikasian dokumen masukan laporan keluhan warga di polrestabes semarang. Penelitian ini dilakukan dengan menggunakan algoritma *cosine similarity* dan *nearest neighbor* dalam proses pengklasifikasikannya. *Cosine similarity* mengukur kesamaan antara dua vector dari sebuah ruang hasil kali yang mengukur cosinus dari sudut diantaranya. *Nearest neighbor* menghitung jarak vektor suatu dokumen dengan vektor bobot dokumen yang lain. Hasil penelitian menunjukkan bahwa akurasi klasifikasi dengan algoritma *cosine similarity* lebih baik dibandingkan dengan algoritma *nearest neighbor* dengan nilai akurasi klasifikasi 93,334% untuk algoritma *cosine similarity* dan 80% untuk algoritma *nearest neighbor*.

Abstract

This research aims to simplify the classification document of report complaints in Polrestabes Semarang. This research used cosine similarity algorithm and nearest neighbor algorithm on classification process. Cosine similarity measure between two vector of an inner product space. Nearest neighbor calculate vector distance of the document to another. The result show that the accuracy of cosine similarity algorithm is better than nearest neighbor algorithm. It has the value 93,334% for cosine similarity algorithm and 80% for nearest neighbor algorithm.

✉ Alamat korespondensi:
Gedung E11 Lantai 2 FT Unnes
Kampus Sekaran, Gunungpati, Semarang, 50229
E-mail: akbardawam1992@gmail.com

ISSN 2252-6811

PENDAHULUAN

Teknologi komunikasi memberikan banyak manfaat dalam memudahkan berkomunikasi pada jarak yang relatif jarak jauh. Dengan adanya kemudahan untuk berkomunikasi ini lembaga kepolisian sebagai lembaga layanan masyarakat memberikan kemudahan kepada masyarakat untuk saling berinteraksi. Lembaga kepolisian memberikan kemudahan kepada masyarakat untuk dapat saling berhubungan dengan adanya layanan Contact Center Polri 110. Dengan layanan ini masyarakat dapat menelepon atau mengirimkan Short Messages Service (SMS) pengaduan selama 24 jam. Diharapkan dengan adanya layanan Contact Center Polri 110 ini kepolisian dapat secara optimal menjalankan tugasnya sebagai institusi yang bertugas untuk mengayomi, dan melayani masyarakat sesuai dengan fungsi kepolisian. Operator system Layanan Contact Center 110 Polrestabes Semarang menjelaskan bahwa laporan masyarakat yang masuk melalui Contact Center 110 akan ditindaklanjuti dan akan diketahui melalui sistem. Setelah adanya laporan, petugas meneruskan laporan kepada Kepolisian Daerah (Polda) dan Kepolisian Resort (Polres) terdekat dengan target minimal 5 sampai 10 menit, petugas bisa tiba di lokasi kejadian. Data laporan yang masuk diklasifikasikan oleh petugas menjadi tiga kategori yakni : divisi informasi, divisi tindak pidana, dan divisi kamtibmas. Dalam observasi yang dilakukan oleh peneliti, laporan yang masuk di Polrestabes Semarang laporan tersebut tertulis dalam buku yang diisi oleh petugas. Dengan cara tersebut dinilai kurang efektif untuk pengolahan data selanjutnya. Untuk mempermudah petugas dalam mengklasifikasikan data laporan yang ada perlu adanya aplikasi yang mengklasifikasikan data laporan yang masuk secara otomatis.

Klasifikasi teks otomatis telah banyak diteliti oleh beberapa peneliti dengan menggunakan berbagai metode. Di antaranya decision tree (Rathee Anju dan Prakash Robin,2013), Naïve Bayesian (Korada al.,2012), K-Nearest Neighbor (D.A. Adeniyi et al., 2014), cosine similarity (Thada Vikas, 2013), serta Self – Organizing Map (Chandra Shekar dan Shoba, 2009).

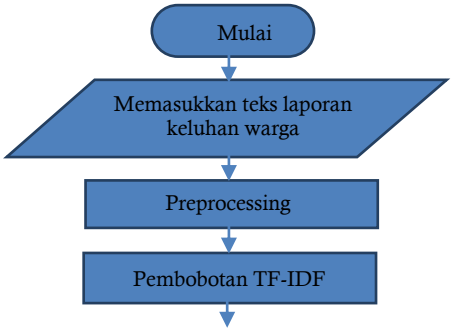
Berdasarkan gambaran yang telah dikemukakan, didapat permasalahan yang diidentifikasi yakni : apakah aplikasi pengklasifikasian otomatis laporan warga di polrestabes semarang dapat diimplementasikan dengan metode Nearest Neighbor dan Cosine Similarity , dan bagaimana tingkat keakuratan klasifikasi teks laporan warga di polrestabes semarang dengan menggunakan metode Nearest Neighbor dan Cosine Similarity.

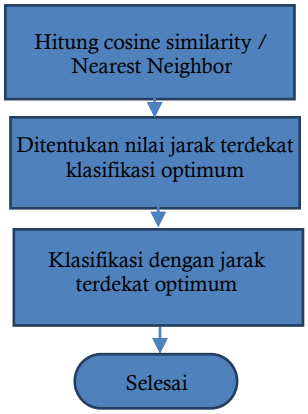
Pemilihan metode Nearest Neighbor dan Cosine Similarity dipilih karena metode tersebut mudah diimplementasikan, cepat dan akurat. Selain itu dalam beberapa penelitian, metode ini menghasilkan hasil yang sangat baik dibanding algoritma yang lain. Maka dari itu banyak peneliti menggunakan metode ini untuk menyelesaikan masalah dalam klasifikasi.

Tujuan dari penelitian ini yaitu untuk merancang aplikasi klasifikasi laporan keluhan warga secara otomatis dengan menggunakan metode Nearest Neighbor dan Cosine Similarity dan untuk mengetahui tingkat akuasi pada aplikasi klasifikasi laporan keluhan warga secara otomatis dengan menggunakan metode Nearest Neighbor dan cosine similarity.

METODE PENELITIAN

Tahapan pertama dalam dalam penelitian ini adalah analisis, untuk mendapatkan aplikasi klasifikasi otomatis yang akan menjadi hasil akhir penelitian ini, identifikaisi kebutuhan perlu dianalisis lebih lanjut. Diagram alur aplikasi klasifikasi otomatis dalam penelitian ini dapat dilihat pada gambar berikut:





Gambar 1. Flowchart Proses Sistem

Setelah tahap analisis kemudian adalah tahap desain. Pada tahap ini dilakukan desain perangkat lunak agar perangkat lunak yang dibuat berjalan sesuai dengan tujuan dibuatnya perangkat lunak ini. Tahapan ini meliputi desain data, desain arsitektur, dan desain interface.

Setelah tahap desain kemudian dilakukan tahap pengkodean. Pada tahap ini dilkukan pembuatan aplikasi menggunakan metode Pemrograman Berorientasi Object (PBO) yang menggunakan aplikasi XAMPP dengan bahasa pemrograman PHP. Pengkodean dilakukan terhadap rancangan – rancangan baik rancangan aplikasi maupun rancangan tampilan. Database yang digunakan menggunakan MySQL untuk menyimpan data. Versi yang digunakan menggunakan MySQL Versi 5.5.16. Untuk text editor dalam pengkodean menggunakan aplikasi Sublime Text 2 Versi 2.0.2.

Pada tahap pengujian dilakukan pengujian aplikasi dan evaluasi hasil klasifikasi otomatis yang dilakukan. Hasil klasifikasi otomatis ini dievaluasi berdasar hasil klasifikasi aplikasi dibandingkan dengan hasil klasifikasi dari ahli dibidangnya.

Pengujian dilakukan dibagi menjadi 2 bagian, yaitu pengujian fungsionalitas aplikasi dan evaluasi hasil klasifikasi.

1. Pengujian fungsionalitas aplikasi
   Pengujian fungsionalitas aplikasi dilakukan untuk memeriksa tidak adanya error pada aplikasi dan aplikasi berjalan sesuai dengan fungsinya dengan baik sesuai dengan tujuan pengujian poin satu.
2. Evaluasi hasil klasifikasi

Untuk mengevaluasi hasil klasifikasi otomatis pada aplikasi dilakukan dengan mencocokkan hasil klasifikasi dari aplikasi dibandingkan dengan data hasil klasifikasi yang telah diklasifikasikan oleh petugas yang ahli dibidangnya. Dari hasil evaluasi klasifikasi aplikasi terhadap data sebenarnya akan dihitung tingkat akurasinya.

HASIL PENELITIAN DAN PEMBAHASAN

Hasil klasifikasi didapatkan setelah melalui tahapan yakni yaitu *preprocessing,* pembobotan TF-IDF, klasifikasi dengan menggunakan algoritma *cosine similarity*, dan algoritma *Nearest Neighbor*.

Pre-processing dalam proses klasifikasi dokumen digunakan untuk membangun sebuah index dari koleksi dokumen. Index adalah himpunan term yang menunjukkan isi atau topik yang dikandung oleh dokumen. Pembuatan inverted index harus melibatkan konsep *linguistic processing* yang bertujuan meng-ekstrak term-term penting dari dokumen yang dipresentasikan sebagai bag-of-words. Ekstraksi term biasanya melibatkan tiga operasi utama, antara lain:

1. *Tokenizing*
   Tokenization adalah tugas memisahkan deretan kata di dalam kalimat, paragraf atau halaman menjadi token atau potongan kata tunggal atau *termmedword*.
2. *Filtering*
   Penghapusan *stop-words. Stop word* didefinisikan sebagai *term* yang tidak berhubungan (*irrelevant*) dengan subyek utama dari *database* meskipun kata tersebut sering kali hadir di dalam dokumen.
3. *Stemming*
   Kata-kata yang muncul di dalam dokumen sering mempunyaibanyak varian morfologik. Karena itu, setiap kata yang bukan *stop-words* direduksi ke *stemmed word* (*term*) yang cocok yaitu kata tersebut di *stem* untuk mendapatkan bentuk akarnya dengan menghilangkan awalan atau akhiran.

Berikut adalah skema pembobotan tf-idf pada bobot term t pada dokumen d (Manning et al.,2009:118).

$$\text{tf} - idf_{t,d} = tf_{t,d} \times idf_t$$

Keterangan :

tf : jumlah kata yang muncul dalam satu kalimat/dokumen

idf :kemunculan kata pada kumpulan kalimat/dokumen

t_i : kata yang akan dihitung bobotnya ke (i=1,2,3…dst) sejumlah kata

d_i :dokumen/kalimat ke (i=1,2,3…dst) sejumlah dokumen

Cosine similarity adalah ukuran kesamaan antara dua vector dari sebuah ruang hasil kali yang mengukur cosinus dari sudut diantaranya. Kosinus 0^o adalah 1, dan kurang dari 1 untuk setiap sudut lainnya. Sehingga orientasinya : dua vektor dengan orientasi yang sama memiliki *cosine similarity* 1, dua vektor pada 90^o memiliki kesamaan 0, dan vektor bertentangan memiliki kesamaan -1, cosine similarity secara khusus digunakan dalam ruang yang positif, berikut adalah persamaan *cosine similarity* (Kanimozhi dan RBalakrishnan J.,2014):

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} * \sqrt{\sum_{i=1}^{n} {B_i}^2}}$$

Keterangan :

A_i : bobot query kata ke (i=1,2,3,…)

B_i : bobot dokumen kata ke (i=1,2,3,…)

Klasifikasi *nearest neighbor* didasarkan pada analogi belajar yakni, dengan membandingkan data tes yang diberikan dengan data pelatihan yang mirip dengannya. Setiap data merepresentasikan titik dalam ruang n-dimensi. Dengan cara ini, semua data pelatihan disimpan dalam ruang pola n-dimensi.

Euclidian distance antara dua titik atau data, maka $X_1 = (x_{11},x_{12},\ldots,x_{1n})$ dan $X_2 = (x_{21},x_{22},\ldots,x_{2n})$ adalah

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

Dari hasil perhitungan dari dua algoritma yaitu algoritma cosine similarity dan nearest neighbor didapatkan hasil :

Tabel 1. Hasil akurasi algoritma cosine similarity

K	Persentasi akurasi	Kesalahan pada saat ujicoba ke-n
1	93,334%	16, 17
2	93,334%	16, 17
3	93,334%	16, 17

Tabel 2. Hasil akurasi algoritma nearest neighbor

K	Persentase akurasi	Kesalahan pada saat ujicoba ke-n
1	80%	7,9,17,19,26,28

Dari tabel 1 dan 2 diketahui nilai persentase terbaik adalah algoritma cosine similarity dengan persentase 93,334% saat berada pada jarak terdekat ke 1,2,3 dari query / data masukan. Sedangkan algoritma nearest neighbor mendapatkan persentase akurasi 80% saat berada pada jarak terdekat ke 1 dari query / data masukan.

Pada jarak terdekat ke 1,2,3 dari data masukan, algoritma cosine similarity mendapatkan persentase akurasi terbesar yaitu 93,334% dengan dua kesalahan klasifikasi yakni pada ujicoba ke 16 dan ujicoba ke 17. Pada dua data masukan tersebut masuk dalam klasifikasi tindak pidana tetapi dalam sistem klasifikasi otomatis dua data masukan tersebut yakni pada ujicoba 16 dan 17 dengan jarak terdekat 1,2,3 seluruhnya terklasifikasi masuk dalam divisi informasi.

Kesalahan klasifikasi pada ujicoba ke 16 diidentifikasi karena kata – kata kunci pada data masukan lebih banyak kesamaannya dengan divisi informasi dibandingakan dengan divisinya sendiri yakni divisi tindak pidana. Diketahui tiga dokumen dengan kesamaan kata kunci terbanyak dengan dokumen masukan yakni dokumen 57 yang masuk pada divisi informasi, dokumen 19 yang masuk pada divisi informasi, dan dokumen 72 yang masuk pada divisi tindak pidana. Pada dokumen ke 57 dan dokumen ke 19 memiliki 2 kesamaan kata dengan data masukan, kemudian pada dokumen ke 72 memiliki 1 kesamaan kata dengan data masukan. Meskipun pada dokumen 57 dan dokumen 19 sama – sama memiliki jumlah kesamaan kata yang sama, namun pada dokumen 57 memiliki nilai idf yang lebih besar dibanding dengan nilai idf dokumen ke 19 sehingga dokumen 57 dianggap memiliki bobot yang lebih besar dibandingkan dengan dokumen ke 19. Sehingga didapatkan hasil klasifikasi dengan jarak terdekat 1,2,dan 3 masuk dalam kategori informasi karena dominasi kata – kata kunci data masukan didominasi pada kategori informasi dibandingkan dengan kategori asli data masukan tersebut yakni kategori tindak pidana.

Kesalahan klasifikasi pada ujicoba ke 17 diidentifikasi karena kata – kata kunci pada data masukan lebih banyak kesamaannya dengan divisi informasi dibandingakan dengan divisinya sendiri yakni divisi tindak pidana.

Diketahui tiga dokumen dengan kesamaan kata kunci terbanyak dengan dokumen masukan yakni dokumen 68 yang masuk pada divisi informasi, dokumen 86 yang masuk pada divisi informasi, dan dokumen 64 yang juga masuk pada divisi informasi. Pada dokumen ke 68, dokumen ke 86 dan pada dokumen ke 64 memiliki 1 kesamaan kata yang sama dengan data masukan. Karena dominasi kata – kata kunci pada data masukan masuk dalam kategori informasi maka kesalahan klasifikasi pada ujicoba ke 17 didapat karena kemiripan kata – kata kunci data masukan lebih banyak didominasi pada divisi informasi dibandingkan dengan divisi tindak pidana.

Dari ketiga nilai persentase akurasi tertinggi yakni pada jarak ke 1,2 dan 3 terdekat dengan data masukan, yang mana mempunyai nilai persentase akurasi yang sama dipilihlah jarak ke 1 terdekat dengan data masukan. Jarak ke 1 dipilih karena pada jarak tersebut memiliki nilai kemiripan tertinggi dan akan mempengaruhi jarak

SIMPULAN

Dari hasil penelitian dengan menggunakan metode Nearest Neighbor dan Cosine Similarity didapatkan hasil akurasi algoritma Cosine Similarity lebih baik dibandingkan dengan algoritma Nearest Neighbor. Hasil klasifikasi yang dilakukan menghasilkan persentase Cosine Similarity sebesar 93,334 % sedangkan persentase 80 % dihasilkan dengan menggunakan algoritma Nearest Neighbor dengan 2 kesalahan klasifikasi pada algoritma Cosine Similarity dan 6 kesalahan klasifikasi dengan menggunakan algoritma Nearest Neighbor. Pada algoritma Cosine Similarity didapatkan kesalahan klasifikasi pada ujicoba ke-16 dan ujicoba ke-17 yang sebenarnya masuk dalam klasifikasi tindak pidana. Kesalahan tersebut disebabkan karena kata kunci pada ujicoba ke-16 dan ke-17 lebih cenderung banyak pada klasifikasi informasi. Dari dua algoritma Nearest Neighbor dan Cosine Similarity dipilihlah algoritma Cosine Similarity sebagai algoritma yang cocok untuk apliasi klasifikasi laporan keluhan warga dipolrestabes semarang karena menghasilkan nilai ketepatan akurasi yang lebih baik dibandingkan dengan algoritma Nearest Neighbor dengan ketepatan akurasi sebesar 93,334%.

DAFTAR PUSTAKA

Chandra Shekar, B.H., dan Shoba, G. 2009. Classification Of Documents Using Kohonen' s Self-Organizing Map. International Journal of Computer Theoryand Engineering. Vol 1, No. 5.

Korada, Naveen Kumar, Kumar, N Sagar Pavan, Deekshitulu, Y V N H. 2012. Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System. International Journal of Information Sciences and Techniques. Vol 2, No. 3. May 2012.

Manning, Christopher D.,Raghavan, Prabhakar, Schütze, Hinrich. 2009. An Introduction to Information Retrieval. Cambridge University Press. :117.

Prabowo, Arif. 2013. Tingkatkan Layanan kepada Masyarakat, Polri Bersinergi dengan Telkom Kelola Contact Center 110 Polri. http://www.telkom.co.id/tingkatkan-polri.html.diakses oktober 2015.

Rathee Anju dan Prakash Robin.2013. Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance. International Journal of Computers & Technology. Volume 4, No. 2. March-April 2013.

Thada, Vikas dan Jaglan, Vivek. 2013. Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm. International Journal of Innovations in Engineering and Technology.: 202-205.

Kanimozhi, R. and J. Balakrishnan, 2014. Cosine similarity based clustering for software testing using prioritization. J. Comput. Eng., 16(1): 75-8