Klasifikasi Surat Laporan Kehilangan Kepolisian Menggunakan Algoritma K – Nearest Neighbor

Ivan Jaya[(1)], Ainul Hizriadi[(2)], Evi Sersanti Purba[(3)]
Program Studi Teknologi Informasi
Universitas Sumatera Utara
Jalan Universitas No. 9A
e-mail : ivanjaya@usu.ac.id

Abstrak

Klasifikasi teks adalah proses pengelompokan dokumen ke dalam kategori atau kelas yang berbeda. Surat laporan kehilangan kepolisian memiliki bermacam – macam jenis, seperti: surat kehilangan Kartu Tanda Penduduk (KTP), surat kehilangan Surat Tanda Tamat Belajar (STTB) dan lain-lain. Klasifikasi surat laporan kehilangan kepolisian masih dilakukan secara manual, karena belum adanya sistem untuk mengklasifikasi surat tersebut. Klasifikasi surat manual memiliki keterbatasan alokasi ruang dan waktu. Untuk menyeselesaikan permasalahan tersebut, penelitian ini menawarkan implementasi algoritma k-Nearest Neighbor untuk mengklasifikasi surat laporan kehilangan kepolisian. Algoritma k-Nearest Neighbor adalah salah satu metode klasifikasi untuk data mining terkhusus text mining. Metode ini bekerja dengan mencari kedekatan jarak suatu data dengan data lain. Pembobotan term dilakukan dengan mencari TF-IDF (Term Frequency-Inverse Document Frequency). Arsip digital surat dibuat melalui proses scanning dan menyimpan isi utama surat dalam file teks. Dalam hal ini surat laporan kehilangan kepolisian digolongkan menjadi tiga kategori utama yaitu kartu, surat, dan sertifikat. Dari hasil pengujian klasifikasi pada 100 isi surat laporan kehilangan kepolisian, algoritma K-Nearest Neighbor dapat menghasilkan ratarata tingkat akurasi 91.75 %.

Kata Kunci** : Klasifikasi, Data Mining, Text Mining, K-Nearest Neighbor, Surat Laporan Kehilangan.**

1. PENDAHULUAN

Instansi kepolisian memiliki tanggung jawab untuk mengelola semua surat laporan yang mereka miliki. Pengelolaan surat tersebut dapat berupa klasifikasi arsip fisik surat berdasarkan tahun atau kategori tertentu. Klasifikasi teks adalah proses pengelompokan dokumen ke dalam kategori – kategori atau kelas – kelas yang berbeda (Joachims, 1997). Surat laporan kehilangan kepolisian memiliki bermacam – macam

DOI: https://doi.org/10.29103/techsi.v10i2.900

jenis, seperti: surat kehilangan Kartu Tanda Penduduk (KTP), surat kehilangan Surat Tanda Tamat Belajar (STTB) dan lain-lain. Hingga saat ini, klasifikasi surat laporan kehilangan kepolisian masih dilakukan secara manual. Hal ini dikarenakan pihak kepolisian belum memiliki sistem terkomputerisasi untuk mengklasifikasikan surat laporan. Namun, klasifikasi surat secara manual memiliki keterbatasan berupa alokasi ruang dan waktu. Klasifikasi surat laporan secara manual memiliki keterbatasan ruang karena media penyimpanan arsip fisik surat yang terbatas, dapat mengalami kerusakan atau gangguan seperti kebakaran maupun kehilangan. Selain itu dari segi waktu, pencarian arsip fisik surat secara manual membutuhkan waktu yang lama. Arsip fisik surat juga sulit dilakukan back-up (pencadangan) secara manual karena besarnya biaya dan tenaga dalam pelaksanaannya. Klasifikasi surat laporan kehilangan kepolisian terkomputerisasi dapat mengatasi keterbatasan alokasi ruang dan waktu.

Fokus utama penelitian ini adalah untuk membuat sistem klasifikasi surat laporan kehilangan. Dalam hal ini surat laporan kehilangan digolongkan menjadi tiga kategori utama yaitu kartu, surat, dan sertifikat. Arsip digital surat akan dibuat untuk mendukung proses pencarian surat melalui sistem. Arsip digital dibuat melalui proses scanning dan menyimpan isi utama surat dalam file teks. File teks digunakan untuk mengekstrak informasi dari dalam surat. Dan informasi digunakan dalam menghasilkan klasifikasi dan melakukan pencarian surat laporan kehilangan.

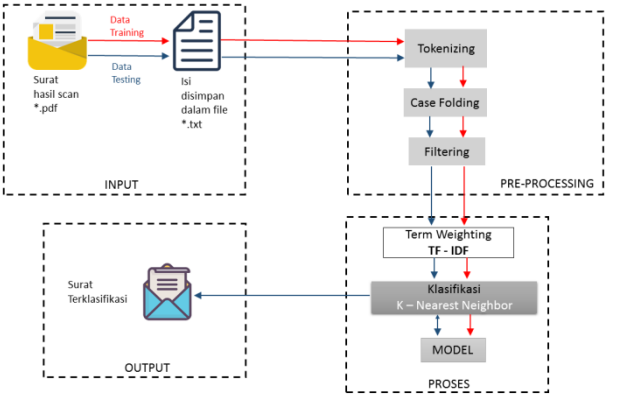
Pada penelitian ini, sistem yang akan dibangun adalah sistem klasifikasi surat laporan kehilangan kepolisian dengan menggunakan algoritma K-Nearest Neighbor. Input dari sistem ini adalah isi surat laporan kehilangan kepolisian dalam beberapa dokumen teks yang akan diproses dengan menggunakan algoritma K-Nearest Neighbor sehingga menghasilkan output berupa klasifikasi surat kedalam kategori tertentu. Dengan pemilihan algoritma K-Nearest Neighbor, input dan output yang akan digunakan, diharapkan algoritma K-Nearest Neighbor akan memberikan hasil klasifikasi yang akurat.

2. METODE PENELITIAN

Metode yang diajukan untuk klasifikasi surat laporan kehilangan ini terdiri dari beberapa langkah. Langkah tersebut yaitu pengumpulan surat laporan kehilangan yang akan digunakan sebagai data *training* dan data *testing*, preprocessing yaitu proses untuk mempersiapkan dan

DOI: https://doi.org/10.29103/techsi.v10i2.900

membersihkan data dari dataset untuk selanjutnya diklasifikasikan. Setelah itu akan dilakukan pembobotan dengan menggunakan metode TF-IDF. Klasifikasi menggunakan algoritma K-Nearest Neighbor akan mengklasifikasi surat laporan kehilangan berdasarkan kategori yang telah ditentukan. Adapun arsitektur umum yang menggambarkan setiap langkah metode yang digunakan dalam penelitian ini ditunjukkan pada Gambar 2.1.



Gambar 2.1. Arsitektur Umum

A. Input

Input yang digunakan pada sistem ini merupakan penggalan isi surat laporan kehilangan kepolisian dalam format *.txt. Dokumen tersebut diinput untuk diolah oleh sistem.

B. Preprocessing

1. *Tokenizing*

Tahap pemotongan teks input menjadi kata, istilah, simbol, tanda baca, atau elemen lain yang memiliki arti yang disebut token disebut *tokenizing*(Vijayarani & Janani, 2016). Pada proses ini, token yang merupakan tanda baca yang dianggap tidak perlu seperti titik (.), koma (,), tanda seru (!), dan lain-lain akan dihapus. Contoh : "Satu lembar kartu ATM Bank BRI atas nama Heru Sudrajat" diubah menjadi token "Satu", "lembar", "kartu", "ATM", "Bank", "BRI", "atas", "nama", "Heru", "Sudrajat".

2. *Case Folding*

Case-folding adalah proses penyamaan case dalam teks. Hal ini disebabkan karena tidak semua teks konsisten dalam penggunaan huruf kapital. Oleh karena itu dilakukan *casefolding* untuk mengkonversi semua teks kedalam suatu bentuk standar (*lowercase*).

DOI: https://doi.org/10.29103/techsi.v10i2.900

Contoh: "Satu lembar kartu ATM" diubah menjadi "satu lembar kartu atm".

3. *Filtering* (*Stopword Removal*)
   Proses yang dilakukan pada tahap ini yaitu menghapus *stop-word*. *Stop-word* adalah kata yang bukan merupakan kata unik dalam suatu artikel atau kata-kata umum yang biasanya selalu ada dalam suatu artikel. Contoh kata yang termasuk stop-word adalah "yang", "dan", "di", "dari", "sedang", "ke", "ini", "oleh"(Tala, 2003). Contoh : "Satu lembar kartu atm dan kartu keluarga" diubah menjadi "satu lembar kartu atm kartu keluarga".

C. Proses Klasifikasi

Langkah-langkah yang dilakukan dalam proses klasifikasi adalah sebagai berikut.

1. *Term Weighting*
   Term Weighting merupakan proses pemberian nilai terhadap setiap term yang terdapat pada tiap dokumen yang telah melewati proses *pre-processing*. Pemberian nilai atau bobot terhadap *term* dalam penelitian ini mengunakan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). TF-IDF merupakan metode yang paling umum digunakan dalam pemberian bobot suatu term. Tujuan dari pembobotan adalah untuk memberikan nilai pada suatu term yang dimana nilai suatu term tersebut akan dijadikan sebagai input pada klasifikasi.
   Tahap pembobotan term dengan TF – IDF adalah sebagai berikut :
   i. Menghitung Nilai TF (*Term Frequency*)
      Setiap term yang ada pada suatu dokumen dihitung nilai kemunculannya. Kemudian nilainya dibagi dengan banyaknya term yang ada pada dokumen tersebut.
   ii. Menghitung DF (*Document Frequency*)
      DF (*Document Frequency*) merupakan banyaknya dokumen dimana suatu term muncul.
   iii. Menghitung IDF (*Inverse Document Frequency*)
      IDF dapat dicari dengan menghitung nilai dari log (N/DF) dengan N = jumlah dokumen.
   iv. Menghitung TF-IDF
      Nilai dari TF-IDF adalah hasil kali antara TF dengan IDF.

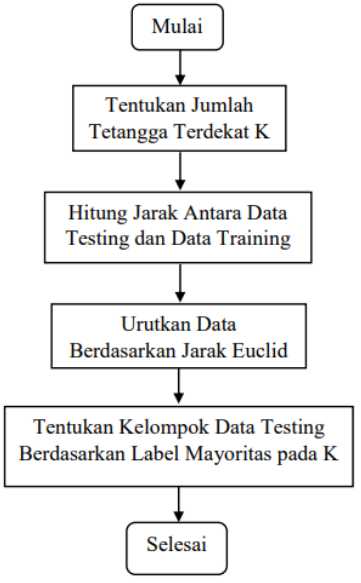
DOI: https://doi.org/10.29103/techsi.v10i2.900

2. *K –Nearest Neighbor*

K-Nearest Neighbor merupakan metode melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Metode ini bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan *training* sample. Apabila sebuah data *testing* yang labelnya tidak diketahui diinputkan, maka K-Nearest Neighbor akan mencari k buah data *training* yang jaraknya paling dekat dengan data *testing* dihitung dihitung dengan cara mengukur jarak antara titik yang merepresentasikan data *training* dengan rumus Euclidian Distance.

Pada fase *training*, algoritma K-Nearest Neighbor hanya melakukan penyimpanan vector-vektor fitur dan klasifikasi data *training* sample. Pada saat fase klasifikasi, semua fitur yang sama dihitung untuk kemudian dilakukan *testing* data yang belum diketahui klasifikasinya (Rivki, 2017).

Penentuan data *training* dan data *testing* harus dilakukan terlebih dahulu sebelum melakukan perhitungan dengan metode K-Nearest Neighbor. Setelah itu, dilakukan proses perhitungan dengan metode KNN seperti Gambar 2.2.



Gambar 2.2. Proses Metode KNN

DOI: https://doi.org/10.29103/techsi.v10i2.900

Klasifikasi data *training* D1, D2, dan D3 secara manual adalah sebagai berikut dengan bobot yang berbeda dimana D1 untuk Kartu, D2 untuk Surat dan D3 untuk Kartu. Terdapat data *testing*yaitu satu lembar surat tanda tamat belajar. Tentukan termasuk kedalam klasifikasi apakah data *testing* tersebut berdasarkan data *training* yang ada, apakah termasuk surat atau kartu.

Langkah penyelesaian:

1. Tentukan parameter K = jumlah tetangga terdekat, pada kasus diatas K = 3.
2. Ambil term data *testing* yang sama dengan D1, D2 atau D3, yaitu : satu, lembar, dan surat. Untuk nilai IDF dari data *testing* diambil dari data *training*.
3. Jumlahkan nilai TF-IDF data *training* D1, D2, dan D3 untuk term : satu, lembar, dan surat.
4. Hitung jarak antara data *testing* dengan semua data *training*.
5. Cari Euclidean Distance dari data *testing* dengan data *training* untuk menghitung jauh dekatnya ketetanggaan.
6. Urutkan hasil kuadrat jarak tersebut secara ascending dan tetapkan tetangga terdekat.
7. Berdasarkan hasil dimana nilai K = 3, diperoleh jarak data *training* dan data *testing* terdekat yaitu 0.016 (D2 / surat). Jadi hasil klasifikasi menggunakan K – Nearest Neighbor data *testing* masuk kedalam klasifikasi surat.

D. Output

Output perancangan sistem ini adalah klasifikasi surat laporan kehilangan dari dokumen yang diinput. Selain itu, surat laporan kehilangan yang merupakan data *training* dapat ditampilkan dalam format *.pdf. Proses klasifikasi telah selesai apabila telah ditentukan kategori dari isi surat yang diinput.Sistem yang dibangun pada penelitian ini adalah sistem untuk menyelesaikan permasalahan klasifikasi surat laporan kehilangan kepolisian. Sebelumnya, sistem klasifikasi surat laporan kehilangan kepolisian belum pernah dilakukan. Algoritma yang diajukan untuk melakukan klasifikasi surat laporan kehilangan kepolisian pada penelitian ini adalah algoritma K-Nearest Neighbor. Algoritma K-Nearest Neighbor ini bertujuan untuk mendapatkan hasil yang tepat dan akurat pada klasifikasi surat laporan kepolisian. Sistem akan mengklasifikasi surat laporan kehilangan kedalam 3 kategori yaitu :

1. Kartu : Sistem akan mengklasifikasi surat laporan sebagai kartu apabila isi dari surat laporan tersebut kehilangan ATM, KTP, Buku

DOI: https://doi.org/10.29103/techsi.v10i2.900

Tabungan, Kartu Tanda Anggota, Kartu Keluarga, Buku Rekening, Kartu BPJS, atau Buku BPKB.

2. Surat : Sistem akan mengklasifikasi surat laporan sebagai surat apabila isi dari surat laporan tersebut kehilangan Surat Keterangan Pindah, Surat Tanda Tamat Belajar (STTB), Surat Asli Perumahan, Slip Penarikan Tabungan, Policy Asuransi, Surat Bukti Gadai, Surat Izin Mengemudi (SIM), Surat Bukti Kredit, STNK, atau Surat Keputusan.
3. Sertifikat : Sistem akan mengklasifikasi surat laporan sebagai sertifikat apabila isi dari surat laporan tersebut kehilangan Akte Lahir, Akte Perkawinan, atau Ijazah.

3. HASIL DAN PEMBAHASAN

Apabila sistem sudah selesai dibangun langkah selanjutnya yang akan dilakukan adalah pengujian sistem. Pengujian sistem ini bertujuan untuk mengetahui apakah sistem berfungsi dengan baik serta mengukur berapa tingkat akurasi algoritma *K-Nearest Neighbor* yang diimplementasikan pada sistem klasifikasi surat laporan kehilangan kepolisian ini. Tingkat akurasi hasil pengujian sistem ini akan dihitung dengan menggunakan rumus *Confusion Matrix* yaitu suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep data mining.

Rumus *Confusion Matrix* melakukan perhitungan terhadap: *recall, precision, dan accuracy* sebagai berikut :

- *Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

$$Recall = \frac{Jumlah\ Surat\ yang\ Dipisahkan\ dengan\ Benar}{Jumlah\ Surat\ Sebenarnya}$$

- *Precission* adalah tingkat ketepatan antara informasi yang yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

$$Recall = \frac{Jumlah\ surat\ yang\ dipisahkan\ dengan\ benar}{Jumlah\ Surat\ yang\ Dipisahkan}$$

- *Accuracy* adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual.

$$Accuracy = \frac{Jumlah\ surat\ yang\ dipisahkan\ dengan\ benar}{Jumlah\ Surat\ Total}$$

DOI: https://doi.org/10.29103/techsi.v10i2.900

Tahap pengujian sistem dilakukan sebanyak 4 kali. Untuk mengetahui akurasi sistem dalam melakukan klasifikasi dilakukan dengan menghitung rata-rata akurasi dari setiap pengujian.Hasil pengujian sistem dapat dilihat pada Tabel 3.1.

Tabel 3.1. Hasil Pengujian Sistem

Pengujian	Jumlah Data	Akurasi
1	25	88%
2	50	92%
3	75	93%
4	100	94%
Rata – rata akurasi =(88%+92%+93%+94%) / 4 = **91.75%**		

Proses pengujiansistemdilakukansebanyakempat kali yaitupengujian1, pengujian2, pengujian3, dan pengujian4denganjumlah data yang berbeda – beda yakni 25, 50, 75 dan 100 data. Setelah dilakukan penghitungan akurasi tiap pengujian, maka ditemukan nilai akurasi rata – rata system sebesar 91.75% dengan nilai rata – rata *recall* sebesar 0.9175 dan *precision* 0.9175.

4. KESIMPULAN

Dari hasil pengujian sistem, diperoleh kesimpulan sebagai berikut:

1. Algoritma K – Nearest Neighbor dapat diterapkan untuk mengklasifikasi surat laporan kehilangan kepolisian dengan tingkat akurasi 91.75 %.
2. Metode text preprocessing yang diajukan mampu mengekstraksi kata kunci (keywords) dari setiap isi surat.
3. Algoritma K – Nearest Neighbor mampu mengelompokkan isi surat yang memiliki kesamaan term.
4. Proses pengujian data *testing* diluar data *training* tergantung pada jarak terdekat serta jumlah anggota kelas yang paling banyak didalam database data *training*.

Saran yang dapat diberikan penulis untuk pengembangan penelitian selanjutnya adalah sebagai berikut:

1. Algoritma *K-Nearest Neighbor* dapat digabungkan dengan algoritma klasifikasi yang lain untuk meningkatkan hasil akurasi.
2. Sistem klasifikasi surat laporan kehilangan ini dapat dikembangkan dalam versi *mobile*.

DOI: https://doi.org/10.29103/techsi.v10i2.900

DAFTAR PUSTAKA

Joachims, T. 1997. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*. Proc of International Conference on Machine Learning (ICML) 1997, USA, Pp. 143-151.

Vijayarani, S., & Janani, R. 2016. *Text Mining: Open Source Tokenization Tools – An Analysis*. Advanced Computational Intelligence: An International Journal (ACII) 3(1), pp. 37-47.

Tala, F.Z. 2003. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. Tesis, Institute for Logic, Languange and Computation: Universiteti van Amsterdam the Netherlands.

Rivki, M. 2017. *Implementasi Algoritma K-Nearest Neighbor dalam Pengklasifikasian Follower Twitter yang Menggunakan Bahasa Indonesia*. Skripsi, Universitas Komputer Indonesia.

DOI: https://doi.org/10.29103/techsi.v10i2.900