

# Pemetaan Tindak Kejahatan Jalanan di Kota Semarang Menggunakan Algoritma *K-Means Clustering*

Widi Astuti<sup>1</sup> dan Djoko Adi Widodo<sup>2</sup>

*Jurusan Teknik Elektro, Fakultas Teknik, Universitas Negeri Semarang, Indonesia*  
*diast824@gmail.com<sup>1</sup>, djokoadiwido@mail.unnes.ac.id<sup>2</sup>*

**Abstrak—** Data kejahatan di Polrestabes Semarang merupakan kumpulan berbagai kasus kejahatan yang kemudian dikelompokkan berdasarkan peraturan yang mengatur yaitu KUHP (Kitab Undang-undang Hukum Pidana) dan Non KUHP, jumlah kejahatan per wilayah polisi sektor, dan jumlah kejahatan per jenis kasus kejahatan. Pengelompokan data dapat memberikan penemuan informasi baru yang sebelumnya tidak diketahui. Penelitian ini bertujuan untuk memetakan data kejahatan jalanan di Kota Semarang periode tahun 2014. Penelitian ini menggunakan metode observasi dan metode studi pustaka. Analisis data menggunakan teknik *text mining* yang terdiri tiga tahapan yaitu *preprocessing*, *representation*, dan *knowledge discovery* (penggalan informasi menggunakan *clustering* algoritma k-means. Selanjutnya dievaluasi menggunakan metode *Entropy*. Hasil penelitian diperoleh 3 *cluster*/kelompok tindak kejahatan jalanan berdasarkan waktu kejadian dan 7 kelompok menurut modus operandi. Berdasarkan nilai *Entropy*, kualitas algoritma k-means dalam pengelompokan waktu kejadian adalah baik karena data yang terkumpul dalam satu kelompok memiliki kemiripan tinggi dibandingkan data pada kelompok lain dan untuk pengelompokan modus operandi adalah cukup baik karena ada satu kelompok yang anggota kelompoknya memiliki tingkat kemiripan yang rendah. Simpulan dari penelitian ini yaitu algoritma k-means dapat digunakan untuk pengelompokan data teks kejahatan jalanan. Saran yang dapat diberikan adalah supaya lebih teliti dalam menentukan jumlah kelompok sehingga data dapat terkelompok dengan baik sesuai kemiripan masing-masing data.

**Kata kunci—** Tindak kejahatan, *Text mining*, *K-Means*, *Entropy*

## I. PENDAHULUAN

Kota Semarang merupakan salah satu kota besar sebagai ibukota propinsi Jawa Tengah. Sebagai sebuah kota besar, kota Semarang memiliki luas wilayah 373,67 km<sup>2</sup> dengan jumlah penduduk 1.268.292 jiwa dan 16 kecamatan. Banyaknya jumlah penduduk menimbulkan berbagai masalah sosial di dalam masyarakat seperti kemiskinan, pengangguran, dan lainnya. Adanya masalah sosial mendorong beberapa orang untuk melakukan tindak kejahatan. Salah satu jenis tindak kejahatan adalah kejahatan jalanan.

Makaampoh dalam menyatakan tindak kekerasan dalam KUHP salah satunya adalah aksi kejahatan jalanan (*Street Crime*) seperti pencurian, pemerasan, pemerkosaan, penganiayaan, tindak kekerasan terhadap orang atau barang, perilaku mabuk di muka umum, yang tentunya dapat mengganggu ketertiban umum serta menimbulkan keresahan di masyarakat [1].

Berdasarkan hasil observasi yang dilakukan di kantor Bidang Operasional (Bin Ops) Reskrim, rincian data kasus kejahatan dikelola oleh Bidang Operasional Reskrim. Pengolahan data kejahatan dilakukan secara manual oleh petugas dengan menggunakan aplikasi pengolah angka Ms. Excel. Adapun atribut data yaitu Nomor Laporan Polisi (No LP), Waktu Kejadian, Tempat Kejadian Perkara (TKP),

Kelurahan, Kecamatan, Tindak Pidana (TP)/Pasal, Kerugian, Modus Operandi (MO), dan Keterangan. Data kasus kejahatan tersebut dikelompokkan sesuai dengan wilayah Polsek (Polisi Sektor) dan menurut KUHP (Kitab Undang-undang Hukum Pidana) atau NON KUHP.

Untuk kasus kejahatan jalanan, seperti pencurian, penculikan, penganiayaan, dan perampasan tidak dicatat secara terpisah. Rincian data kejahatan jalanan masih dicatat bersamaan dengan kasus kejahatan lainnya. Adapun data kejahatan jalanan yang diolah lebih lanjut antara lain jumlah kasus pencurian yang terjadi menurut wilayah Polsek.

Berdasarkan latar belakang di atas, maka dirumuskan permasalahan sebagai berikut yaitu bagaimana kualitas algoritma *k-means clustering* dalam memetakan tindak kejahatan jalanan meliputi penganiayaan, pencurian, penjambratan, perampokan, dan pemerasan di kota Semarang pada tahun 2014.

Menurut Linoff dan Berry, clustering adalah proses pengelompokan satu set data objek menjadi beberapa kelompok atau klaster sehingga objek dalam sebuah klaster memiliki kemiripan yang tinggi satu sama lain, tetapi sangat berbeda dengan objek dalam kelompok lainnya. Salah satu algoritma yang sering digunakan adalah algoritma k-means [2].

## II. METODE

Penelitian ini menggunakan metode observasi dan metode studi pustaka. Observasi dilakukan di Polrestabes Semarang bagian Bidang Operasional (Bon Ops) Kasat Reskrim. Dari observasi diperoleh data kejahatan selama periode tahun 2014 sebanyak 799 data.

Analisis data menggunakan teknik *text mining* yang terdiri tiga tahapan yaitu *preprocessing* atau praproses (*tokenization, case folding, filtering, stemming*), *representation* (penyusunan vektor menggunakan pembobotan *TF/Term Frequency*), dan *knowledge discovery* (penggalian informasi menggunakan *clustering*) dan dievaluasi menggunakan metode *Entropy* untuk mengetahui kualitas algoritma *k-means* dalam melakukan pengelompokan.

Praproses merupakan tahap yang dilakukan untuk mentransformasi data tekstual yang bersifat tidak terstruktur menjadi model yang terstruktur. Terdapat beberapa tahap proses yang dilakukan dalam tahap praproses yaitu proses *tokenization, case folding, filtering, dan stemming*. *Tokenization* merupakan proses pemisahan teks menjadi potongan kata yang disebut *token*. *Case folding* merupakan proses pengubahan huruf dalam dokumen menjadi satu bentuk, misalnya huruf kapital menjadi huruf kecil dan sebaliknya. *Filtering* merupakan proses mengambil kata-kata yang penting dari hasil token. *Stemming* merupakan proses untuk menemukan akar kata atau kata dasar dari sebuah kata.

Representasi dokumen digunakan untuk mengubah versi teks dokumen ke bentuk vektor. Model vektor yang digunakan adalah skema pembobotan *term frequency* (TF). Dalam merepresentasi digunakan dua notasi yaitu 0 dan 1. Notasi 0 berarti bahwa tidak ada term tertentu yang terdapat dalam suatu dokumen, sedangkan notasi 1 menandakan bahwa ada suatu term tertentu yang berada dalam suatu dokumen. Representasi vektor dengan pembobotan TF ditunjukkan pada Tabel I.

TABEL I. REPRESENTASI VEKTOR DENGAN PEMBOBOTAN TF

|   | Jumat | Kamis | Minggu | Rabu |
|---|-------|-------|--------|------|
| 1 | 0     | 0     | 0      | 1    |
| 2 | 0     | 0     | 0      | 1    |
| 3 | 0     | 0     | 0      | 1    |
| 4 | 0     | 1     | 0      | 0    |
| 5 | 0     | 1     | 0      | 0    |

Penggalian informasi menggunakan *clustering* pada data berupa teks disebut *document clustering*. *Document clustering* atau pengelompokan dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunakan metode klaster (*clustering*) data mining [3].

Salah satu algoritma yang sering digunakan adalah algoritma *k-means*. *K-Means* merupakan salah satu metode

*cluster analysis* non hirarki yang berusaha untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* atau kelompok objek berdasarkan karakteristiknya, sehingga objek yang mempunyai karakteristik yang sama dikelompokkan dalam satu *cluster* yang sama dan objek yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam *cluster* yang lain [4].

Karena pada prinsipnya analisis klaster adalah untuk mengelompokkan objek berdasarkan pada kemiripan, maka diperlukan suatu metode untuk mengukur kemiripan atau perbedaan antarobjek. Pengukuran kemiripan menggunakan *Euclidean Distance* dimana pengukuran jarak dilakukan dengan menghitung akar kuadrat dari penjumlahan kuadrat perbedaan dari nilai masing-masing variabel [5].

Evaluasi digunakan untuk mengukur kualitas dari algoritma yang digunakan. Dalam penelitian ini evaluasi yang digunakan untuk mengukur kualitas adalah *Entropy*. Pengukuran *Entropy* adalah untuk mengukur kemurnian dari *cluster* yang dihasilkan dengan memperhatikan pada kategori ada. Nilai *Entropy* yang lebih kecil menghasilkan *cluster* yang lebih bagus kualitasnya. Keakuratan hasil *Entropy* berada pada jangkauan 0 – 1 dimana semakin kecil hasil *Entropy*-nya, maka kualitas *cluster* semakin baik [6]. Besar kecilnya nilai *Entropy* berkaitan dengan keberagaman informasi dalam suatu *cluster*. Dalam Teori Informasi *Entropy* menunjukkan ketidakaturan sesuatu (dalam hal ini informasi). Semakin tinggi nilai *Entropy* sebuah *message* (informasi) berarti semakin banyak simbol yang muncul dalam informasi tersebut.

## III. HASIL DAN PEMBAHASAN

Berdasarkan hasil pengelompokan data menggunakan algoritma *k-means clustering*, didapatkan hasil *clustering* seperti ditunjukkan pada Tabel II. Dari Tabel II dapat diketahui nilai *Entropy* untuk masing-masing *cluster* atribut waktu kejadian. *Cluster*/kelompok 1 sebesar 0,177091, *cluster* 2 sebesar 0,165072, dan *cluster* 3 sebesar 0,185785 dengan jumlah keseluruhan sebesar 0,527947 dan rata-rata sebesar 0,175982. Berdasarkan nilai *Entropy* yang dihasilkan maka dapat disimpulkan bahwa ketiga *cluster* memiliki kualitas yang baik karena nilai *Entropy* mendekati 0. Dengan demikian, anggota dalam satu *cluster* memiliki banyak kemiripan, tetapi dengan objek pada *cluster* lain memiliki banyak perbedaan.

TABEL II. HASIL EVALUASI ENTROPY PADA WAKTU KEJADIAN

| Cluster   | Jumlah data | Entropy  |
|-----------|-------------|----------|
| 1         | 261         | 0,177091 |
| 2         | 245         | 0,165072 |
| 3         | 273         | 0,185785 |
| Total     | 779         | 0,527947 |
| Rata-rata |             | 0,175982 |

TABEL III. HASIL EVALUASI ENTROPY PADA MODUS OPERANDI

| Cluster   | Jumlah data | Entropy  |
|-----------|-------------|----------|
| 1         | 127         | 0,069551 |
| 2         | 2           | 0,000057 |
| 3         | 27          | 0,005827 |
| 4         | 414         | 0,257583 |
| 5         | 44          | 0,013227 |
| 6         | 126         | 0,068758 |
| 7         | 39          | 0,010828 |
| Total     | 779         | 0,425831 |
| Rata-rata |             | 0,060833 |

Tabel III menunjukkan nilai *Entropy* untuk masing-masing *cluster* atribut modus operandi. *Cluster*/kelompok 1 sebesar 0,069551, *cluster* 2 sebesar 0,000057, *cluster* 3 sebesar 0,005827, *cluster* 4 sebesar 0,257583, *cluster* 5 sebesar 0,013227, *cluster* 6 sebesar 0,068758, dan *cluster* 7 sebesar 0,010828 dengan jumlah keseluruhan sebesar 0,425831 dan rata-rata sebesar 0,060833. *Cluster* yang paling baik adalah *cluster* 2 dengan nilai *Entropy* 0,000057, kemudian *cluster* 3 dengan nilai *Entropy* 0,005827 dan *cluster* yang paling tidak baik adalah *cluster* 4 dengan nilai *Entropy* 0,257583. Dengan demikian, anggota dalam *cluster* 2 memiliki banyak kemiripan, dibandingkan dengan anggota dalam *cluster* 4.

#### IV. SIMPULAN DAN SARAN

Berdasarkan hasil penelitian dan pembahasan tersebut diatas dapat disimpulkan bahwa kualitas algoritma k-means untuk pengelompokan data teks adalah baik. Hal ini diperoleh dari nilai *Entropy* yang mendekati angka 0 dengan rata-rata nilai *Entropy* untuk waktu kejadian adalah 0,175982 dan untuk modus operandi adalah 0,060833. Dengan demikian algoritma K-means dapat digunakan untuk pengelompokan data kejahatan jalanan.

#### REFERENSI

- [1] Fadlina. 2014. Data Mining untuk Analisa Tingkat Kejahatan Jalanan dengan Algoritma Association Rule Metode Apriori. Majalah Ilmiah 3(1): 144-154.
- [2] Megawati, C. 2015. Analisis Aspirasi dan Pengaduan di Situs LAPOR! Dengan Menggunakan Text Mining. Skripsi. Program Studi Teknik Industri Universitas Indonesia. Depok.
- [3] Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. 2012. Partical Text Mining and Statistical Analysis for Non-Structured Text Data Application. Oxford: Elsevier.
- [4] Ediyanto, Muhasah N. M., dan Neva S. 2013. Pengklasifikasian Karakteristik dengan Metode K-Means Cluster Analysis. Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster) 02(2): 133-136.
- [5] Suliyanto. 2005. Analisis Data dalam Aplikasi Pemasaran. Bogor: Ghalia Indonesia
- [6] Langgeni, D. P., ZK. A. Baizal, dan Yanuar F. A. W. 2010. Clustering Artikel Berita Berbahasa Indonesia Menggunakan Unsupervised Feature Selection. Seminar Nasional Informatika 2010 UPN "Veteran" Yogyakarta. 22 Mei 2010: D-1 – D-10.