Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 5, Mei 2019, hlm. 4434-4438 http://j-ptiik.ub.ac.id

Klasifikasi Dokumen pada Laporan Kepolisian dengan Menggunakan Metode BM25 dan *Improved K-Nearest Neighbor* (IKNN)

Ardhimas Ilham Bagus Pranata[1], Indriati[2], Marji[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]ardhimasilham@ymail.com, [2]indriati.tif@ub.ac.id, [3]marji@ub.ac.id

Abstrak

Kepolisian Negara Republik Indonesia merupakan salah satu penegak hukum yang ada di Negara Kesatuan Republik Indonesia. Salah satu tugas Kepolisian Indonesia adalah memberikan pelayanan kepada masyarakat. Pengaduan tindak kejahatan merupakan salah satu bentuk pelayanan kepada masyarakat yang ditawarkan oleh Kepolisian. Tahapan setelah laporan tindak kejahatan diterima oleh kepolisian adalah diterbitkannya surat untuk proses penyelidikan dan proses penyidikan. Namun dalam kurun waktu satu bulan, kepolisian khususnya Polres Kota Malang mengalami kesulitan untuk mengklasifikasi seluruh laporan kejahatan yang diterima. Oleh sebab itu dibutuhkan sistem yang dapat membantu kepolisian dalam mengklasifikasi laporan kejahatan kedalam tiga kasus kejahatan yang terdiri dari penganiayaan, pencurian dan penipuan. Proses dalam penelitian ini adalah dengan melakukan *pre-processing text* yang selanjutnya akan dihitung bobot nilai tf, df, idf serta dilanjutkan dengan proses klasifikasi. Dalam penelitian ini klasifikasi dilakukan dengan menggunakan metode BM25 dan *Improved K-Nearest Neighbor* (IKNN). Dari hasil pengujian dengan *k-fold cross validation* didapatkan rata-rata nilai tertinggi *precision*=0,953373, *recall*=0,931382, *f-measure*=0,938122 dan *accuracy*=0,956795 pada nilai k=15.

Kata kunci: *klasifikasi, pre-processing text, BM25, improved k-nearest neighbor, k-fold cross validation.*

Abstract

The National Police of the Republic of Indonesia is one of the law enforcers in the Unitary State of the Republic of Indonesia. One of the tasks of the Indonesian Police is to provide services to the community. Accusation of crime is one form of service to the community offered by the Police. Crime can happen to anyone no matter an employee, a student or others. The stage after the report of a crime is received by the police is the issuance of investigation. However, within one month the police had difficulty classifying every police report that have been accepted especially Polres Kota Malang. Therefore a system for helping the police to classified a accusation of crime into three cases are persecution, stealing, and fraud is needed. Process in this study is by doing a pre-processing text which the next stage is counting the weight of tf, df, and idf and continue to classification. In this study classification do by using BM25 and Improved K-Nearest Neighbor Methods (IKNN). The results of the k-fold cross validation test, the highest average value of precision=0,953373, *recall*=0,931382, *f-measure*=0,938122 *and accuracy*=0,956795 *at the value of k = 15.*

Keywords: *classification, pre-processing text, BM25, improved k-nearest neighbor, k-fold cross validation.*

1. PENDAHULUAN

Kepolisian Negara Republik Indonesia merupakan salah satu penegak hukum yang ada di Negara Kesatuan Republik Indonesia. Sebagai penegak hukum kepolisian memiliki tugas yang antara lain adalah pemeliharaan keamanan dan ketertiban masyarakat, menegakkan hukum, memberikan pengayoman pada masyarakat, memberikan perlindungan serta memberikan pelayanan pada masyarakat (Undang-undang Nomor 2 Tahun 2002). Dalam penulisan penelitian kali tugas kepolisian yang diangkat adalah tentang memberikan pelayanan kepada masyarakat. Salah satu bentuk pelayanan kepada masyarakat yang ditawarkan oleh kepolisian adalah pengaduan tindak kejahatan. Di Indonesia sendiri khususnya di Kota Malang,

dalam kurun waktu satu bulan sebanyak 300 laporan kejahatan yang diterima oleh Polres Kota Malang.

Kejahatan dapat terjadi kepada siapa saja tidak peduli apakah korban merupakan seorang pegawai, pelajar, mahasiswa atau yang lainnya. Dalam penelitian ini terdapat tiga jenis kejahatan yang digunakan yaitu penganiayaan, pencurian, dan penipuan. Sebagai pelayan masyarakat, kepolisian memiliki tahapan dalam memproses laporan kejahatan yang diberikan oleh masyarakat. Tahapan setelah laporan diklasifikasi adalah diterbitkan surat proses penyelidikan dan proses penyidikan. Namun Polres Kota Malang mengalami kesulitan dalam mengelompokkan semua laporan yang diterima dalam waktu satu bulan. Oleh sebab itu penelitian ini dilakukan dengan menggunakan *Improved K-Nearest Neighbor* sebagai metode klasifikasi dan BM25 sebagai metode untuk pemeringkatan dokumen.

Improved K-Nearest Neighbor merupakan salah satu metode dari klasifikasi yang paling umum digunakan yaitu *K-Nearest Neighbor* (KNN). Kelebihan *Improved K-Nearest Neighbor* terdapat di penggunaan nilai *k* yang berdasarkan pada nilai jumlah data latih dari tiap kelas. *Improved K-Nearest Neighbor* juga menghasilkan hasil klasifikasi yang lebih stabil bila dibandingkan dengan *K-Nearest Neighbor* (Li, Yu and Lu, 2003).

BM25 merupakan sebuah metode yang digunakan untuk menghitung tingkat kemiripan (*similarity) dari query* dengan kumpulan dokumen dengan menggunakan tiga faktor penentu untuk menghitung *score* BM25 yang terdiri dari *term frequency* (tf); *Inverse Document Frequency* (IDF); dan rata-rata panjang dokumen (Tjandra and Widiasri, 2016).

Maka berdasarkan penjabaran masalah diatas, maka penulis ingin menerapkan kedua metode di atas untuk klasifikasi dokumen laporan kepolisian dengan menggunakan metode BM25 dan *Improved K-Nearest Neighbor.*

2. DASAR TEORI

2.1. Kepolisian Negara Republik Indonesia

Kepolisian negara republik Indonesia (POLRI) merupakan salah satu penegak hukum yang ada di Indonesia. Dalam struktur organisasinya polri bertanggung jawab secara langsung kepada Presiden Indonesia. Polri memiliki tugas pada seluruh wilayah Indonesia. Tugas dari polri adalah sebagai memelihara keamanan dan ketertiban masyarakat; menegakkan hukum; memberikan perlindungan, pengayoman, dan pelayanan pada masyarakat (Undang-undang Nomor 2 Tahun 2002).

2.2. Sistem Temu Kembali Informasi

Sistem temu kembali informasi merupakan bagian dari pengolahan pada teks yang digunakan untuk mendapatkan informasi yang serupa atau sesuai dengan apa yang dicari oleh penggunanya. Dalam sistem temu kembali informasi terdapat proses pencarian; penyimpanan; serta pemeliharaan informasi. Informasi yang didapatkan dapat berupa teks, audio, gambar, dan video atau dapat berupa media yang lain (Dgz and Ferdinandus, 2015)

2.3. *Pre-processing text*

Pre-processing text merupakan proses awal yang perlu dilakukan ketika query atau dokumen akan diberikan bobot nilai. *Pre-processing text* bertujuan untuk merubah data yang tidak terstruktur menjadi terstruktur. Adapula tujuan lain dari *pre-processing text* adalah menghilangkan *noisy* pada data sehingga nantinya didapatkan hasil yang lebih optimal. Hasil dari *pre-processing text* adalah didapatkan term dalam bentuk kata dasarnya (Feldman and Sanger, 2007).

Dalam *pre-processing text* ada beberapa tahapan yang dilakukan antara lain adalah *case folding*, *tokenizing*, *filtering* dan *stemming*.

2.3.1. Case Folding

Case folding merupakan tahapan pada query dan dokumen yang diberikan oleh *user* diubah dari kata-kata mengandung huruf besar (*uppercase*) menjadi ke dalam bentuk huruf kecil (*lowercase*) (Puspitasari et al., 2017).

2.3.2. Tokenizing

Tokenizing merupakan proses untuk merubah *query* dan dokumen yang mulanya berbentuk kalimat diubah menjadi satuan kata. Dalam tokenizing dilakukan proses menghilangkan karakter selain huruh a sampai z (Puspitasari et al., 2017).

2.3.3. Filtering

Filtering merupakan tahapan dimana *query* dan dokumen dihilangkan kata-kata yang dianggap tidak penting, dengan berdasarkan

pada *stopword list* (Puspitasari et al., 2017).

2.3.4. Stemming

Stemming merupakan tahap yang terjadi pada *query* dan dokumen proses dalam stemming adalah pengembalian kata berimbuhan menjadi ke dalam kata dasarnya. Dalam teks berbahasa Indonesia stemmer menghilangkan mengubah kata menjadi kata dasar (Adriani et al., 2007).

2.4. BM25

Metode BM25 merupakan salah satu metode yang dapat digunakan untuk menghitung tingkat kemiripan dokumen (*similarity*) dari *query* yang diberikan terhadap kumpulan dokumen yang tersedia. Dalam menghitung tingkat kemiripan dokumen, BM25 menggunakan 3 faktor antara lain adalah *term frequency* (tf); *Inverse document frequency* (IDF); rata-rata panjang dokumen (Russell and Norvig, 2013). Perhitungan dari *score* BM25 menggunakan Persamaan (1).

$$Score = \sum_{i=1}^{N} IDF(q_i) \frac{TF_{(q_i,d_j)} \cdot (k+1)}{TF_{(q_i,d_j)} + k(1-b+b\frac{|d_j|}{L})} \quad (1)$$

Serta untuk menghitung nilai IDF menggunakan Persamaan (2).

$$IDF(q_i) = \log \frac{N - DF_{(q_i)} + 0{,}5}{DF_{(d_j)} + 0{,}5} \quad (2)$$

Keterangan:
$IDF(q_i)$: nilai *invers df* query *i*
$TF(q_i d_j)$:*Term frequency* q_i pada dokumen *j*
L : Rata-rata panjang dokumen
$|d_j|$: Panjang isi dokumen j
N : Jumlah dokumen
$k : 1.2 \leq k \leq 2$
$b : 0 \leq b \leq 1$

2.4. *Improved K-Nearest Neighbor*

Improved K-Nearest Neighbor merupakan salah satu metode dalam klasifikasi yang memiliki kelebihan dibanding dengan metode *K-Nearest Neighbor* (KNN). Kelebihan tersebut terdapat pada metode penggunaan nilai *k* dengan berdasarkan pada jumlah data latih yang ada pada tiap kelas (Li, Yu and Lu, 2003).

Dalam metode *improved k-nearest neighbor* terdapat proses untuk menghitung nilai k_1 baru. Perhitungan nilai k_1 baru atau n menggunakan persamaan (3).

$$n = \left[\frac{k * N(C_m)}{Maks\{n(C_m)|j=1..N_c\}}\right] \quad (3)$$

Serta untuk menghitung nilai *improved k-nearest neighbor* dengan menggunakan Persamaan (4).

$$P(x, C_m) = argsMaks_m \frac{sim(x,d_j)y(d_j,C_m)}{sim(x,d_j)} \quad (4)$$

Keterangan:
n : nilai k baru
k : nilai k yang ditetapkan
$N(C_m)$: Jumlah data latih pada kategori m
$Maks\{n(C_m)|j = 1 \dots N_c\}$: nilai max pada seluruh kelas
$P(x, C_m)$: Peluang dokumen x anggota dari kelas m
$sim(x, d_j)$: Kemiripan dokumen x dengan dokumen j
$y(d_j, C_m)$: Fungsi atribut 1 apabila d_j termasuk C_m, 0 jika tidak.

2.5. Evaluasi

Evaluasi dilakukan untuk mengetahui seberapa besar tingkat kinerja dari metode yang digunakan. Dalam penelitian ini evaluasi dilakukan dengan menggunakan *confusion matrix.* Pada *confusion matrix* terdapat beberapa kriteria yang digunakan untuk menghitung nilai *f-measure*, *recall*, *precision*, dan *accuracy* yang antara lain adalah *true positive* (tp), *true negative* (tn), *false positive* (fp), *dan false negative* (fn) (Powers, 2007).

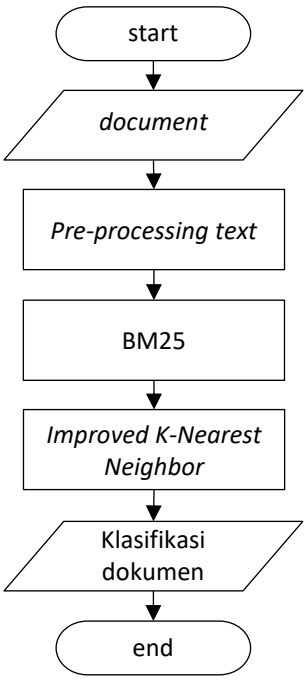
3. METODE PENELITIAN

3.1. Pengumpulan Data

Dalam penelitian ini pengumpulan data dilakukan dengan cara observasi. Observasi dilakukan di Polres Kota Malang pada bidang Reserse Kriminal untuk mengambil data sebanyak 100 data dalam kurun waktu 2 minggu atau sekitar 12 hari serta untuk mengamati serta mencari tahu secara langsung pada lokasi yang memiliki hubungan dengan penelitian. Dalam penelitian ini data yang digunakan merupakan data laporan kepolisian mulai dari bulan Februari 2018 sampai dengan Juni 2018.

3.2. Perancangan Sistem

Dalam perancangan sistem akan dijelaskan secara singkat bagaimana sistem atau alur kerja secara umum. Alur kerja sistem ditunjukkan pada Gambar 1.



Gambar 1. Alur kerja sistem

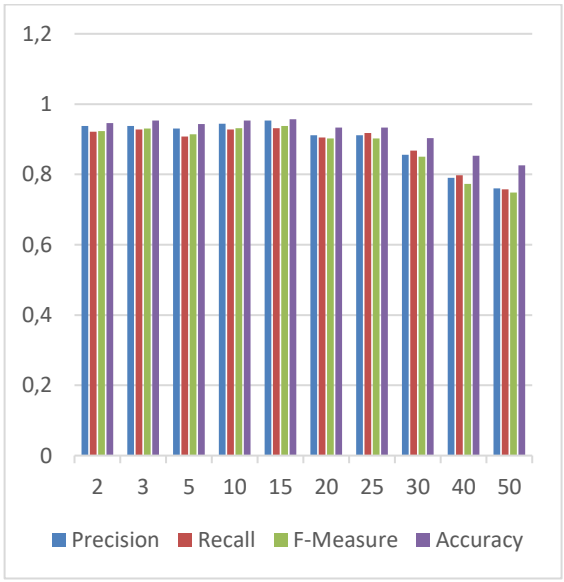
Berdasarkan Gambar 1, pada tahap pertama akan dilakukan *pre-processing text* pada data uji dan data latih. Kemudian setelah dilakukan *pre-processing text* tahap selanjutnya adalah dihitung nilai *score* BM25 dengan menggunakan 3 faktor yang antara lain adalah *term frequency* (tf), *inverse document frequency* (idf), dan panjang rata-rata document. Setelah mendapatkan *score* BM25 maka tahapan berikutnya adalah diklasifikasi dengan menggunakan metode *improved k-nearest neighbor*.

4. PENGUJIAN DAN ANALISIS

Pengujian dilakukan dengan menggunakan k-*fold cross validation*. Proses pengujian dalam penelitian ini adalah, pengujian dilakukan sebanyak 4 pengujian. Dimana dalam setiap pengujian menggunakan sebanyak 25 data sebagai data latih serta 75 data sebagai data uji dengan rincian 21 data merupakan data penganiayaan, 36 data merupakan data pencurian, dan 18 data merupakan data penipuan. Hasil rata-rata nilai *precision*, *recall*, *f-measure*, dan *accuracy* dari 4 pengujian ditunjukkan pada Tabel 1.

Tabel 1. Hasil rata-rata pengujian *k-fold cross validation*

k	Precision	Recall	F-measure	Accuracy
2	0,9375	0,921429	0,923042	0,945797
3	0,9375	0,92791	0,930307	0,953333
5	0,930556	0,907407	0,913765	0,943333
10	0,944444	0,92791	0,93166	0,953333
15	0,953373	0,931382	0,938122	0,956795
20	0,911706	0,90463	0,90214	0,933333
25	0,911706	0,91746	0,90251	0,933333
30	0,856151	0,867454	0,850169	0,903333
40	0,790675	0,797756	0,77275	0,853333
50	0,759921	0,757858	0,748164	0,826087



Gambar 2. Grafik rata-rata pengujian

Berdasarkan pada pengujian yang telah dilakukan, maka nilai precision, recall, f-measure, dan accuracy tertinggi sebesar 0.953373, 0.931382, 0.938122, dan 0.956795 pada nilai *k*=15 dan terandah pada nilai *k*=50 dengan nilai precision= 0.759921, recal=0.757858, f-measure=0.748164 dan accuracy=0.826087. Hal tersebut terjadi karena pada nilai *k*=15 akan menghasilkan nilai k_1 baru dalam *improved k-nearest neighbor* sebesar 21 untuk kelas penganiayaan, 36 untuk kelas pencurian, dan 18 untuk kelas penipuan. Sedangkan pada nilai *k*=50 akan menghasilkan nilai *k* baru sebesar 29 untuk kelas penganiayaan, 50 untuk kelas pencurian, dan 25 untuk kelas penipuan.

Perhatikan grafik pada Gambar 2. Dari nilai k_1 baru tersebut akan berpengaruh terhadap ketetanggaan terdekat dengan data uji yang diberikan. Semakin besar nilai k_1 baru yang dihasilkan maka akan berpengaruh pula pada perhitungan nilai probabilitas dari *improved k-nearest neighbor*, karena dengan nilai ketetanggaan yang terlalu besar maka akan menghitung semakin banyak pula nilai dokumen negatif dari *score* BM25. Nilai dokumen negatif merupakan nilai *score* BM25 yang dimiliki oleh *term* yang terdapat pada hampir semua dokumen yang digunakan.

5. PENUTUP

5.1. Kesimpulan dan Saran

Berdasarkan penelitian yang telah dilakukan, maka dalam penelitian ini dapat di ambil kesimpulan serta saran sebagai berikut.

Metode BM25 dan Improved K-Nearest Neighbor dapat digunakan dalam melakukan klasifikasi dokumen laporan kepolisian dengan menghasilkan rata-rata *precision*=0,953373, *recall*=0,931382, *f-measure*=0,938122, serta *accuracy*=0,956795 ketika nilai k_1 yang diberikan sebesar 15.

Serta saran yang diberikan penulis bagi penelitian berikutnya adalah gunakan metode *Improved* BM25 atau *Modified* IDF BM25 agar didapatkan hasil yang lebih optimal, dan gunakan dokumen dengan term yang bersifat *universal* atau term yang pasti dimiliki oleh banyak dokumen yang tersedia.

DAFTAR PUSTAKA

Adriani, M., Asian, J., Nazief, B. and Williams, H., 2007. Stemming Indonesian : *A confix-stripping approach . Stemming Indonesian* : *A Confi x-Stripping Approach.* Universitas Indonesia, Depok.

Dgz, S. and Ferdinandus, F., 2015. Sistem *Information Retrieval* Layanan Kesehatan untuk Berobat dengan Metode *Vector Space Model*(VSM) Berbasis WEBGIS. Sekolah Tinggi Informatika & Komputer Indonesia, Malang.

Feldman, R. and Sanger, J., 2007. *The Text Mining Handbook*. Cambrige University Press, New York.

Li, B., Yu, S. and Lu, Q., 2003. An Improved k-Nearest Neighbor Algorithm for Text Categorization. *Institute of Computational Linguistics*. Peking University, Beijing.

Powers, D.M.W., 2007. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. School of Informatics and Engineering. Flinders University*, Adelaide, Australia.

Puspitasari, A.A., Santoso, E., Yusuf, T., Harjoko, A., Studi, P. Ilmu, J. Teknis, P. Produksi, P., Tanah, K. and Hijau, K., 2017. Klasifikasi Dokumen Tumbuhan Obat Menggunakan Metode Improved K-Nearest Neighbour. *Journal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(2),113–123.

Russell, S. and Norvig, P., 2013. *Artificial Intelligence A Modern Approach*. [online] *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*.

Tjandra, E. and Widiasri, M., 2016. Sistem Repositori Tugas Akhir Mahasiswa dengan Fungsi Peringkat Okapi BM25. *Journal of Information Systems Engineering and Business Intelligence*, [online] 2(2), 88-94. Universitas Airlangga, Surabaya.

Undang-undang Nomor 2 Tahun 2002, Tentang Kepolisian Negara Republik Indonesia. Indonesia, Jakarta.