Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 6, Juni 2019, hlm. 5697-5702 http://j-ptiik.ub.ac.id

Klasifikasi Emosi Lirik Lagu menggunakan *Improved K-Nearest Neighbor* dengan Seleksi Fitur dan BM25

Febrina Sarito Sinaga[1], Indriati[2], Bayu Rahayudi[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]febrinasinaga.fs@gmail.com, [2]indriati.tif@ub.ac.id, [3]ubay1@ub.ac.id

Abstrak

Emosi merupakan reaksi atau perasaan seseorang terhadap suatu keadaan. Emosi bersifat temporer yang dapat muncul oleh stimulus karena orang-orang sekitar dan lingkungannya. Salah satu contoh lingkungan yang dapat merangsang emosi seseorang adalah berasal dari lagu yang didengarkan. Lirik lagu merupakan bagian yang berperan membangun emosi. Pemilihan kata-kata yang tepat diperlukan untuk menciptakan emosi yang sesuai. Dalam hal ini klasifikasi emosi pada lirik lagu akan dilakukan proses pengklasifikasian menggunakan beberapa metode yaitu Improved K-Nearest Neighbor, BM25 dan seleksi fitur. Proses klasifikasi ini terdapat beberapa tahapan didalamnya yaitu tahapan melakukan pre-processing dokumen, tahapan perhitungan score BM25 dan pemeringkatan pada setiap dokumen, dan tahapan klasifikasi dengan algoritme yang digunakan adalah Improved K-Nearest Neighbor. Pengujian untuk pengklasifikasian ini dilakukan sebanyak 6 kali pengujain *k-fold* dan menggunakan confusion matrix . Penelitian ini jumlah data latih yang digunakan sebanyak 100 dokumen, dan jumlah data uji yang digunakan sebanyak 20 dokumen. Pada semua pengujian yang telah dilakukan diperoleh hasil rata-rata yang cukup baik saat nilai k = 55 dengan hasil f-measure sebesar 0,6693, recall sebesar 0,6582, dan precision sebesar 0,7427.

Kata kunci: *klasifikasi, seleksi Fitur, BM25, improved KNN*

Abstract

Emotions is a person's reaction or feeling into a situation. Emotion is temporary that can occurred by a stimulus because of some people around and the environment. One of example an environment that can trigger someone's emotion is from the song being listened to. Song lyrics are the parts that can build emotions. Choosing the right words for lyrics are very important because it will create the right emotion. In this case the emotional classification of song lyricis will be done classifying process using several methods are Improved K-Nearest Neighbor, BM25 and feature selection. The proses of classification have seome stages, which is the stage of pre-processing documents, stages of calculation the BM25 score and sorting document, and the classification stage with using the algorithm is Improved K-Nearest Neighbor. The testing for classifications was done uses 6 times K-fold and use the confusion matrix. This research is the amount of training data used by 100 documents, and testing data used by 20 testing documents. In the all the tests have done obtained the best average results when the value K = 55 with a result of f-measure is 0.6693, recall is 0.6582, and precision is 0.7427.

Keywords: *classification, Feature Selection, BM25, Improved KNN*

1. PENDAHULUAN

Musik merupakan suatu hiburan, seni ataupun aktivitas yang dilakukan dan didengarkan dari manusia dengan dengan perlibatan suara yang teratur. Proses musik memunculkan suatu emosi yang memiliki keterkaitan dengan otak manusia. Musik dapat digunakan untuk penggambaran ekspresi atau perasaan seperti contoh orang mendengar music untuk melepaskan emosi terhadap perasaannya sekarang, music juga dapat melepaskan stress, menghibur diri dan pengambaran ekspresi yang lainnya (Sakti, 2010).

Lirik lagu adalah bagian penting untuk menggambarkan suatu perasaan atau emosi seseorang yang dapat digunakan dalam

penentuan emosi. Lirik lagu merupakan suatu penyampaian perasaan seseorang secara tidak langsung terhadap apa yang dialami, dirasakan, dilihat, ataupun didengarnya. Menentukan emosi berdasarkan lagu biasanya dilakukan secara manual oleh manusia, misalnya seseorang sedang mendengarkan lagu yang mempunyai tempo yang ceria sehingga dia menyimpulkan emosi yang disampaikan dari lagu tersebut adalah emosi yang bahagia, akan tetapi pengarang lagu tersebut belum tentu menggambarkan isi dari lagu tersebut memiliki perasaan bahagia kemungkinan perasaan yang ingin disampaikan adalah perasaan tersipu atau malu. Keterbatasan tersebut mendorong diciptakannya klasifikasi emosi lirik lagu agar lebih efektif jika seseorang yang sedang ingin mendengarkan lagu sesuai dengan emosi yang diinginkan.

Semakin banyaknya lagu semakin banyak jenis emosi yang muncul maka diperlukan pengklasifikasian secara otomatis untuk menyelesaikan masalah tersebut. Pengklasifikasian emosi pada lirik lagu berdasarkan fitur-fitur yang terdapat didalam lagu tersebut. Fitur merupakan suatu ciri data dalam suatu objek yang membedakannya terhadap objek yang lain dan dapat digunakan dalam pengklasifikasian lirik lagu yaitu terdapat pada kata-kata yang sering muncul maupun yang jarang muncul didalam lirik lagu tersebut. Metode pengklasifikasian suatu objek sangat banyak misalnya pengklasfikasian sederhana yaitu *Naïve Bayes*, *K-Means*, *KNN*, *Extra Trees*, dan metode lainnya yang pastinya menggunakan metode yang berbeda menghasilkan akurasi yang berbeda juga.

Penggunaan metode *KNN* adalah bagian dari banyaknya metode pengklasifikasian yang mudah diterapkan juga cukup sederhana. Metode ini mengklasifikasikan pada objek yang yang memiliki jarak terdekat dengan objek yang lain pada suatu data (Tan, 2006). Metode ini juga mempunyai kelemahan karena hanya *k-values* yang sama yang digunakan untuk semua kategori kelas tanpa memperhatikan jumlah data yang belum tentu meliki kemiripan pada tiap ketgorinya. Tetapi dari kelemahan dari metode K-Nearest Neighbor sangat dapat diatasi dengan metode Improved K-Nearest Neighbor (Baoli, 2003). Contoh penelitian sebelumnya dengan meode *BM25* telah dilakukan oleh Syarif Hidayatullah (2014) untuk mendeteksi kemiripan isi dokumen yang diduga plagiarisme menghasil presentase kemiripan masing-masing 65,32%, 37,01% dan 10,24%.

Berdasarkan permasalahan dari uraian sebelumnya penulis memutuskan mengangkat topik penelitian yaitu penglasifikasian emosi lirik lagu menggunakan *Improved K-Nearest Neighbor yang mengambil* nilai dari tetangga terdekatnya, menggunakan seleksi fitur agar dapat mengetahui perbedaan dari penggunaan seleksi fitur dan tanpa seleksi fitur, dan *BM25* untuk melakukan perankingan dokumen.

2. LANDASAN KEPUSTAKAAN

2.1. Emosi Lrik Lagu

Emosi adalah perasaan atau keadaan pikiran naluri alami yang dialami oleh seseorang, suasana hati atau *mood* serta hubungan dengan orang lain. Lirik lagu adalah rangkaian kata-kata yang berisi ekspresi seseorang mengenai apa yang telah dirasakan, dilihat dan didengar oleh pencipta lagu tersebut. Pencipta lagu memainkan melodi, permainan kata-kata dan bahasa untuk meningkatkan daya tarik dan suatu khas terhadap lirik. Sehingga paduan rangkaian kata yang dipadukan dengan melodi dan notasi musik membuat seseorang yang mendengarkannya juga ikut merasakan apa yang dirasakan oleh pengarang lagu tersebut (Sari, Yuita Arum. 2012).

2.2. Klasfikasi *Text*

Pengklasifikasikan teks merupakan teknik mengelompokkan data-data dan dikelompokkan ke dalam sebuah kategori tertentu. Klasifikasi teks merupakan salah satu teknik dalam teks mining. Setiap dokumen memiliki kategori spesifik yang akan menemu-kembalikan *(retrieve)* dokumen ketika kueri yang diberikan sesuai. Teknik klasifikasi teks yang digunakan adalah teknik yang diklasifikasikan oleh seorang pakar terlebihi dahulu. Hal ini dilakukan agar pengelompokan dokumen tersebut akurat dan memiliki karateristik yang sesuai.

2.3. *Pre-Processing*

Tahapan ini merupakan bagian tahapan didalam *text mining* untuk mengurangi dan melakukan peringkasan waktu komputasi karena hanya berfokus terhadap *term* yang penting saja. Pada tahap pre-processing ini dibagi ke dalam 3 tahapan, yakni tokenisasi, penghilangan *stopword* dan *stemming*.

2.3.1 *Tokenisasi*

Pada tahap pre-processing, tahapan tokenisasi adalah proses mencecah dokumen menjadi kata perkata terpisah. Proses tokenisasi dilakukan berdasarkan karakter spasi. Hal tersebut dilakukan karena spasi merupakan pemisah antar kata dalam sebuah kalimat.

2.3.2 *Filterisasi*

Pada tahapan ini adalah tahap penghapusan kata apabila terdapat kata yang tidak diperlukan pada hasil yang telah dilakukan pada tokenisasi. digunakan serta tidak penting disebut *stopwordStopword* biasanya memiliki frekuensi yang tinggi dalam kalimat, apabila dihilangkan tidak akan mempengaruhi makna dari kalimat tersebut seperti kata "yang", "ke", "di","dari".

2.3.3 *Stemming*

Pada tahapan ini yaitu mengubah token menjadi kata dasar. Stemming dilakukan setelah proses filterisasi. Contoh dari proses stemming adalah kata "menggunakan" diubah menjadi kata dasar. Maka awalan me-, -kan akan dihapus.

2.4. Seleksi Fitur

Tahap yang digunakan pada tahapan ini adalah menghapus variabel atau objek yang dianggap berlebihan atau tidak memiliki relevansi pada data. Untuk melakukan seleksi fitur dilakukan penyimpanan kata dengan nilai yang paling tinggi. Seleksi fitur juga mengurasi waktu ruang pencarian dengan menggunakan kebutuhan komputasi yang lebih rendah (Wijaya & Muslim, 2016).

2.5. Improved K-Nearest Neighbor

Tahapan ini adalah tahapan untuk melakukan proses klasifikasi dengan menggunakan *K-vaalue.*. Nilai K tersebut adalah berdasarkan dari banyaknya yang tetangga terdekat untuk tiap kategori. Sedangkan pada K-Nearest Neighbor menggunakan 1 nilai k pada keseluruhan kategori (Baoli, Shiwen dan Qin, 2003). Penentuan nilai k-values yang bagus diperlukan agar hasil akurasi yang didapatkan tinggi. Hasil akurasi didapatkan setelah dilakukannya proses klasifikasi pada data latih dan data uji. Pehitungan *Improved KNN* menggunakan rumus:

$$v = \left(\frac{a * D\ (C_m)}{maks\ \{D(c_m)|j=1 \ldots\ldots D_c\}}\right) \quad (1)$$

Keterangan:
v : new k values
a : k-values awal
$D\ (C_m)$: jumlah data latih pada kategori m
$maks\ \{D(c_m)|j = 1 \ldots\ldots D_c\}$: untuk semua kategori jumlah dokumen latih yang terbanyak

Sejumlah n dokumen dipilih adalah top n dokumen yang memiliki kemiripan paling besar di setiap kategori. Selanjutnya adalah menghitung peluang dokumen uji X termasuk ke dalam kategori m dengan rumus:

$$p(B, c_m) = \arg max_m \frac{\Sigma d_j \in top\ n\ kNN(C_n) sim(b,d_j) y(d_j.c_n)}{\Sigma d_j \in top\ n\ kNN(C_n) sim(x,d_j)} \quad (2)$$

Penjelasan
$p(B, c_n)$: Probabilitias pada *document K* menjadi bagian kategori c_n
$sim(K, d_j)$: Jumlah *document* pada suatu kelas pada data latih d_j
$top\ n\ kNN$: top *n* tetangga
$y(d_j.c_n)$: fungsi atribut yang memenuhi suatu kategori tertentu, akan bernilai 1 apabila dokumen latih d_j masuk kedalam anggota c_m jika tidak maka akan bernilai 0

2.6. BM25

BM25 adalah metode untuk membuat peringkat dokumen berdasarkan permintaan penarian yang diberikan pada kerangka pengambilan probabilistic. Perhitungan BM25 menggunakan rumus:

$$\text{Score (D,Q)} = \sum_{k=0}^{n} IDF\ (Xi) . \frac{ft(x_i,D).(N_1+1)}{ft(X_i,D)+N_1.(1-C+C.\frac{|d|}{\text{avgdl}})} \quad (3)$$

Keterangan:
$f(Xi, D)$: *Frequency term qi* dalam *document*
k_1 : Konstanta frekuensi term dalam dokumen 1,2 ≤ k1 ≤ 2
C : Konstanta panjang dokumen = 0≤b≤1
D : Panjang dokumen D
Avgdl : rata – rata panjang dokumen secara keseluruhan

Pada Rumus diatas terdapat nilai IDF. Nilai tersebut dapat dihitung dengan rumus dibawah ini:

$$IDF(xi) = log \frac{N - a(Xi) + 0.5)}{ft(X_i,D)+N_1.(1-b+b.\frac{|d|}{\text{avgdl}})} \quad (4)$$

Keterangan:
N : Jumlah dokumen data latih
$a(x_i)$: Kemunculan term yang mirip dari

dokumen uji dan dokumen latih

2.7. Evaluasi

Evaluasi termasuk proses melakukan analisis program dan memberikan penilaian dari hasil yang dikeluarkan program. Hal ini dilakukan untuk memberikan keefektifan yang baik dari program yang digunakan (Patton, 1987). Confussion Matrix adalah salah satu jenis evaluasi dalam pengklasifikasi kelas. Tabel perhitungan *confusion matrix* terdapat pada Tabel 1.

Tabel 1. Confusion Matrix

	Nilai Salah (Pakar)	Nilai Benar (Pakar)
Nilai Salah (Prediksi)	True Negative	False Negative
Nilai Benar (prediksi)	False Positive	True Positive

Penjelasan :
TN : Hasil data prediksi benar saat hasil pakar *negative*
FN : Hasil data prediksi salah saat hasil pakar positive
FP : Hasil data prediksi salah saat hasil pakar negative
TP : Hasil data prediksi benar saat hasil pakar *positive*

Selanjutnya dilakukan perhitungan mencari hasil evaluasi mengukur kinerja sistem menggunakan rumus-rumus diantaranya:
Recall yaitu tentang membandingkan jumlah positif dari sistem berdasarkan kesesuaian dari pakar. Perhitungan *recall* dapat ditampilkan pada Rumus (5).

$$Rc = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

Precision yaitu tentang jumlah kebenaran hasil positif dari sistem berdasarkan kesesuaian dari pakar. Menghitung *precision* ditampilkan pada Rumus (6).

$$Ps = \frac{True\ Positive}{False\ Positive + True\ Positive} \quad (6)$$

Penjelasan *F-Measure* yaitu tentang mengukur hasil keseimbangan timbal-balik *recall* dengan *precision.* Rumus dari *F-Measure* dapat ditampilkan pada Rumus (7).

$$FM = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (7)$$

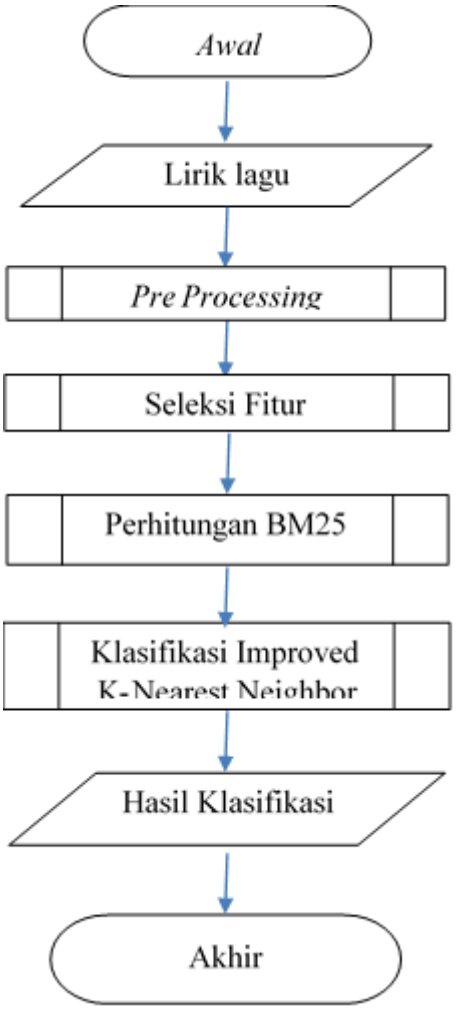
3. METODOLOGI

3.1 Perancangan Algoritme

Perancangan Algoritme merupakan suatu gambaran langkah kerja dalam pembuatan suatu sistem keseluruhan yang akan digunakan untuk tahapan implementasi. Didalam perancangan Algoritme tersebut membahas tentang tahapan kerja sistem dimulai dari inputan, proses dan menghasilkan keluaran.

3.2 Implementasi Alur Sistem

Alur Sistem adalah gambaran umum proses keseluruhan dari sistem yang terdiri dari *pre-processing*, perhitungan score BM25, perhitungan seleksi fitur, *dan* penggunaan *Improved KNN* serta pengujian untuk mendapatkan hasil klasfikasi. Data yang digunakan yaitu lirik lagu dengan pembagian menjadi data uji dan data latih. Gambaran alur sistem ditampilkan pada Gambar 1.



Gambar 1. Alur Sistem

Semua data akan diproses terlebih dahulu dalam *pre-processing*. Selanjutnya dilakukan perhitungan *TF* dan IDF, setelah itu dilakukan seleksi fitur yaitu mengambil term yang hanya muncul lebih dari 1 kali.

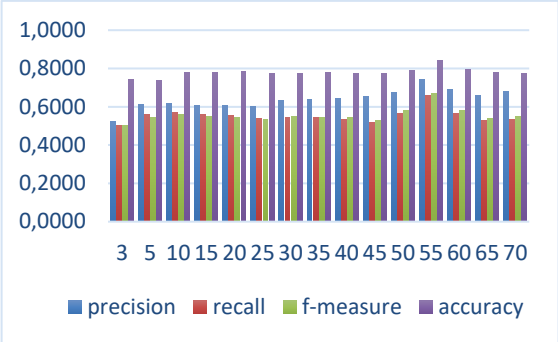
Selanjutnya adalah perhitungan score BM25 dan perankingan score BM25. Setelah mendapatkan hasil perangkingan *score* tersebut dilakukan proses pengklasifikasian berdasarkan *k-values* awal menggunakan Improved KNN.

4. PENGUJIAN DAN ANALISIS

Klasifikasi emosi lirik lagu data yang digunakan berjumalah 120 dengan pembagian jumlah dokumen latih sebanyak 100 dan jumlah dokumen uji sebanyak 20. Dilakukan pengujian sebanyak 6 K-Fold agar mendapatkan k-values baru, *accuracy*, *f-measure*, *recall* dan *precision.* Setiap pengujian menggunakan 20 data uji dengan jumlah pembagian 7 kelas bahagia, 6 kelas sedih, 3 kelas marah, dan 4 kelas malu. *f-measure*, *recall*, dan *precision.* yang didapatkan berdasarkan proses perhitungan *new k-values* tiap nilai *k* dari seluruh hasil pengujian terdapat dalam Tabel 2 dan Gambar 2.

Tabel 2. Rata-Rata Pengujian *F-Measure*, *Recal* dan *Precison*

(k-values)	*F-Measure*	*Recall*	*Precision*
3	0,5008	0,5020	0,5239
5	0,5425	0,5595	0,6129
10	0,5590	0,5724	0,6147
15	0,5479	0,5575	0,6073
20	0,5452	0,5531	0,6043
25	0,5350	0,5382	0,6013
30	0,5470	0,5461	0,6304
35	0,5454	0,5427	0,6363
40	0,5431	0,5347	0,6448
45	0,5296	0,5174	0,6547
50	0,5786	0,5625	0,6731
55	0,6693	0,6582	0,7427
60	0,5791	0,5640	0,6883
65	0,5376	0,5258	0,6602
70	0,5486	0,5327	0,6815



Gambar 2. Grafik Hasil Rata-Rata *Recall*, *Precision* dan *F-Measure*

Berdasarkan seluruh proses dari awal hingga pengujian maka *factor k-values* awal terhadap new k-values berpengaruh kepada hasil pengklasifikasian algortime *Improved K-Nearest Neighbor* dikarenakan *k-values* yang kecil ataupun besar yang ditentukan tersebut sangat berpengaruh terhadap hasil klasifikasi. Dimulai saat *k-values*=3 hingga *k-values* berada saat bernilai 55, hasil *f-measure, recall dan precision* turun naik hasil yang dialami dengan rentang yang kecil karena memiliki nilai tetangga yang sedikit dibandingkan dengan *k-values* saat berada sebesar 60 hingga *k-values* diposisi 70, karena memiliki jumlah yang tetangga yang cukup banyak maka nilai *recall*, *f-measure*, dan *precision*, mengalami penurunan. Terjadinya penurunan juga terjadi dikarenakan banyak perbedaan hasil klasifikasi dari pakar dengan hasil dari sistem terutama pada kelas malu dan marah

Kelas marah dan malu banyak terdapat kesalahan dalam proses klasifikasi ini dikarenakan memiliki sedikit data latih jika dibandingkan dari kelas bahagia dan sedih. Terjadi banyak banyak kesalahan pada kelas sedih saat proses pengklasifikasian dikarenakan data pada kelas sedih mempunyai banyak isi yang mirip dengan data bahagia, contohnya pada kata lupa, menangis, cemburu, galau, akhir, kenang yang seharusnya menjadi term di kelas sedih tetapi masuk kedalam term bahagia mengakibatkan hal itu mengganggu proses klasifikasi.

Maka, pada penelitian ini dapat disimpulkan bahwa seleksi fitur memengaruhi hasil klasifikasi karena hanya mengambil kemunculan nilai term lebih dari satu. Perhitungan dari BM25 sangat cocok untuk melakukan pengurutan data-data berdasarkan pada jumlah data dengan tingkat relevansinya yang sangat baik hingga jumlah data dengan tingkat relevansi yang sangat rendah atau tidak

relevansi sama sekali. Penggunaan algoritme Improved KNN dapat berjalan dengan sangat baik apabila menggunakan data latih yang sangat banyak serta pemilihan k-value terhadap nilai k-baru tersebut mempengaruhi hasil klasifikasi.

5. KESIMPULAN

Dalam pengklasifikasian ini menghasilkan kesimpulan diantaranya, hasil yang didapatkan dari pengujian klasifikasi emosi lirik lagu ini yaitu saat *k-value*=55 menghasilkan *f-measure* tertinggi 0,6693 dan hasil terendah saat *k-value*=3 menghasilkan *f-measure* sebesar 0,5008.

Untuk semua pengujian yang dilakukan hasil nilai k awal dan *new* k-value memengaruhi naik turunnya *f-measure*, *Recall*, dan *Precison* dari k=3 sampai k=70. Berdasarkan banyaknya data latih yang digunakan sangat memengaruhi hasil prediksi pada pengujian untuk menentukan klasifikasi.

Adapun tahapan-tahapan yang dapat dilakukan dalam penelitian klasifikasi emosi lirik lagu ini terdiri dari beberapa proses diantaranya perhitungan dengan metode BM25, preprocessing text, seleksi fitur, dan tahapan pengklasifikasian dengan Improved K-Nearest Neighbor

6. DAFTAR PUSTAKA

Baoli, Siwen, Qin. 2003. *An Improved K-Nearest Neighbour Algorithm for Text Categorization. To Apear in the Procedings of the 20th International Conference of Computer Processing of Oriental Languag.*, Shenyang, China.

Burhan, Muhammad Yasser., Marji & Muflikhah Lailil. 2013. *Pengelompokkan Emosi Berdsarkan Lirik Lagu Menggunakan Metode Iterative Dichotomizer Tree*. S1. Universitas Brawijaya

David M.W. 2007. *Evaluation: From Precision, Recall and F-Measure To ROC, Informedness, Markedness & Correlation [e-journal]*. 2(1), 37-63.

Ferdinan, A. 2018 . Klasifikasi Emosi Pada Lirik Lagu Menggunakan Metode K-Nearest Neighbor. Tersedia pada:<https://jopenlibrary.telkomuniversity.ac.id> [Diakses 16 Januari 2019].

Hidayatullah, Syarif. 2014 Source Detection pada Kasus Plagiarisme dokumen Menggunakan Metode Biword Winnowing dan Retrieval Berbasis Okapi BM25

Patton, M.Q., 1987. How to Use Qualitative Methods in Evaluation. 2. [online] Los Angeles: SAGE Publications. Tersedia pada: <https://books.google.co.id/books?id=0co1ESOVJHkC>[Diakses 29 Februari 2019

Tala, F.Z., 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Master of Logic Project Institute for Logic, Language and Computation Universiteit van Amsterdam