Jurnal Teknologi Informatika dan Terapan Vol. 04, No 01, Januari – Juni 2017 ISSN: 235-838X

KLASIFIKASI LAGU DAERAH INDONESIA BERDASARKAN LIRIK MENGGUNAKAN METODE TF-IDF DAN NAÏVE BAYES

Pujo Hari Saputro [1], Michael Aristian [2], Dyah Listianing Tyas [3]
pujoharisaputro@gmail.com, michael.aristian@gmail.com, dyahlistyaningtyas@gmail.com
Fakultas Teknologi Informasi
Universitas Hasyim Asy'ari Tebuireng Jombang
2017

Abstrak

This research is one of the efforts to classify regional songs in Indonesia by region. This classification is expected to be one way to know and map the local songs in Indonesia so that the nation of Indonesia can recognize its own culture. The amount of data used in this study is 90 songs from various regions. Regional songs will be classified by region. This research will test extraction method of feature of term frequency - inverse document frequency (TF-IDF) and snaïve bayes as its classification method. This research proves that TF-IDF extraction method and classification with naïve bayes can be used to classify the lyrics of regional songs based on the area of the song with 73.4% accuracy on the set of West Indonesia and East Indonesia.

Keywords: Classification of lyrics, regional songs, TF-IDF, naïve bayes.

1. PENDAHULUAN

Indonesia merupakan bangsa yang memiliki berbagai kebudayaan dan kekayaan seni. Salah satu bentuk karya seni yang ada pada semua daerah adalah lagu. Semua daerah di Indonesia dari sabang sampai merauke memiliki lagu daerah yang unik dan memiliki ciri khas sendiri-sendiri tiap daerah. Di Indonesia sendiri menurut harian KOMPAS pada tahun 2010 jumlah lagu daerah ada 485 lagu.

Lagu daerah / rakyat adalah salah satu alat komunikasi yang penting bagi masyarakat (S.Daudu, 2009), sejak tahun 1940 sudah dilakukan banyak upaya untuk merancang system untuk mengkategorikan melodi namun setelah beberapa dekade belum ada teori dan metode yang kuat,dan system klasifikasi lagu daerah belum muncul hingga banyak lembaga warisan budaya memberikan prioritas yang tinggi untuk digitalisasi dan unlocking dari music daerah (Frans Wiering, 2009).

Fokus kami disini adalah mengklasifikasikan lagu daerah berdasarkan liriknya,sebab ada level semantic yang hanya melekat pada lirik dan tidak bisa dideteksi pada audio (Rauber, 2011). Pengklasifikasian lagu biasanya dilakukan dengan ektraksi musik atau ekstraksi lirik atau keduanya (Matt McVicar, 2011) (Min-Yen Kan, 2008). Kami berpendapat bahwa aspek linguistik dalam suatu lagu daerah sangat erat hubungannya dengan asal daerah tersebut,selain itu kami menggunakan lirik sebagai bahan utama sebab bagian dari informasi semantic lagu berada pada liriknya (Cyril Laurier, 2008). Lirik lagu sering menjadi peran penentu dalam kesamaan persepsi 2 lagu serta masuknya mereka dalam jenis tertentu (Robert Neumayer, 2007). Pendekatan yang kami lakukan adalah mengklasifikasikan lagu berdasar lirik dengan menggunakan metode tf*idf untuk pembobotan kata dan menggunakan multi-nominal naive bayes.

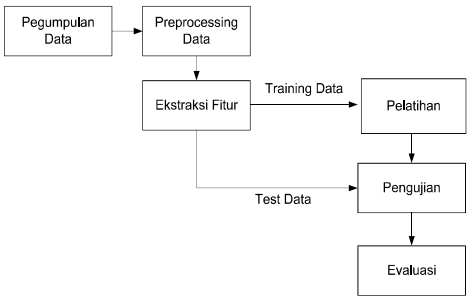
2. TINJAUAN PUSTAKA

Penelitian sebelumya yang pernah dilakukan oleh (Dewi, 2011) dengan judul Incorporating Global Information In a Folk Song Classification System ,dimana dalam tesisnya beliau mengelompokkan lagu lagu rakyat eropa dan menklasifikasikannya secara otomatis, beliau sadar

begitu pentingnya music rakyat atau lagu daerah itu dikelompokkan dan juga ada sebuah media yang bisa digunakan untuk menyimpan directory lyric berdasarkan daerah asalnya. Pendekatan yang digunakan dalam pengklasifikasian music daerah ini adalah pendekatan linguistic dan juga pendekatan N-Gram, beliau berpendapat bahwa kedua metode tersebut lebih efektif dalam pengklasifikasian sebuah lirik. Metode penarikan keputusan yang lain juga pernah digunakan dalam pengklasifikasian lirik seperti : *Hidden Markov Model* Oleh Chai & Vercoe, (2001), *Support Vector machine and time decomposition* Oleh Lidy et al., (2009), juga Naive Bayes, *Logistic Regression, k-nearest neighbors dan decision tree* Oleh Hillewaere, Manderick & Conklin (2009) .Kemudian (Xiaohu, 2009) dalam penelitian mereka yang berjudul *Liric Text Mining in Music Mood Classification* ditulis bahwa Lirik sebuah music sangatlah unik dan membutuhkan teknik preprocessing khusus untuk mengklasifikasikannya, jadi untuk mengklasifikasikan sebuah lagu baik itu menurut mood ,genre ataupun asal daerah lagu tersebut memang membutukan beberapa metode penarikan keputusan yang maksimal.

3. METODE PENELITIAN

Dalam pengklasifikasian lagu, pada umumnya terdiri dari berapa tahapan yaitu: pengumpulan data, pra proses data, ektraksi fitur/ciri, clustering, klasifikasi, pengujian, dan evaluasi.



Gambar 12. Diagram Kerja

3.1. Pengumpulan Data

Pengumpulan data merupakan proses awal dalam membuat suatu sistem pengklasifikasian. Data yang dikumpulkan bisa berasal dari berbagai sumber bisa dari koleksi pribadi, ataupun hasil unduhan dari internet. Saat ini sudah ada beberapa situs yang telah menyediakan suatu data set yang digunakan untuk menguji sistem pengklasifikasian yang dibuat contoh: last.fm, 7digital, crayoon, dan code.soundsoftware.ac.uk/projects-/emotion recognition. (Song, Dixon, & Pearce, 2012).

3.2. Pra-Proses Data

Format data yang dikumpulkan bisa berbeda, data yang berbeda susunan atau format tersebut tidak boleh langsung dijadikan sebagai data unutk penelitian. Setelah data dikumpulkan maka akan dilakukan proses preprocessing. Tahapan ini digunakan untuk menyeragamkan format data yang akan di gunakan dalam pemodelan ataupun pengujian.

3.3. Ektraksi Fitur dan Ciri

3.3.1. Metode TF*IDF

Pembobotan yang digunakan adalah metode tf*idf. Tf*IDF merupakan salah satu metode yang digunakan untuk mengekstraksi ciri dari suatu text. Metode ini merupakan gabungan antara metode term frequency (tf) dengan metode inverse document frequency(idf).

Term frequency(tf) merupakan suatu metode yang digunakan untuk mencari bobot suatu kata dalam dokumen kunci di setiap kategori dan mencari kata kunci yang hampir mirip dengan kategori yang tersedia. Dimana lirik akan disusun (ditransformasikan) menjadi vector serta vector ada pada setiap dimensi adalah kata yang ada pada dokumen lirik lagu. (Zaanen & Kanters, 2010) (Xing Wang, 2011)

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

Semakin banyak suatu kata dalam suatu kelas terkandung dalam suatu dokumen, maka skor tf akan menjadi lebih tinggi pada kelas tersebut. Kelemahan metode ini adalah jika suatu kata muncul pada semua dokumen, maka tidak bisa dipastikan bahwa suatu kata tersebut memiliki makna khusus yang relevan.

Untuk mengatasi masalah tersebut, salah satu cara yang umum digunakan adalah dengan menggunakan metode inverse document frequency(idf). Metode idf ini digunakan untuk menghitung jumlah dokumen yang mengandung kata yang dimaksud, kemudian dibagi dengan total dokumen yang ada.

$$idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (2)$$

Dengan tujuan mendapatkan pembobotan yang sesuai untuk tiap term dalam tiap dokumen, maka dikalukan kombinasi antara metode tf dengan metode idf.

$$tf * idf_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

dengan penjelasan sebagai berikut

a. Bobot paling tinggi jika kata *t* sering terdapat pada jumlah dokumen yang kecil.
b. Bobot agak rendah jika kata *t* sering terdapat pada jumlah dokumen yang besar atau kata *t* jarang terdapat pada jumlah dokumen yang kecil.
c. Bobot paling rendah jika kata *t* terdapat pada setiap dokumen.

3.4. Pengklasifikasian

3.4.1. Naïve Bayes

Naive Bayes merupakan sebuah algoritma pembelajaran yang berbasis pada teori Bayes dengan menggunakan asumsi yang kuat(naive). Teori Bayes merupakan suatu teori tentang mencari suatu probabilitas sesuatu berdasarkan data yang telah ada sebelumnya. Methode ini juga bisa digunakan untuk mengklasifikasikan musik berdasarkan data yang telah dilatih sebelumnya.

Inti dari Naive Bayes adalah mencari probabilitas terbesar suatu musik dalam suatu kategori. Formula Bayes dapat dituliskan sebagai berikut:

$$P(c|d) = \frac{P(c) \times P(d|c)}{P(d)} \quad (4)$$

Dimana $P(c|d)$ adalah probabilitas kelas c setelah d dimasukkan pada kelas c , $P(c)$ probabilitas kelas c sebelumnya, $P(d|c)$ probabilitas d pada kelas c , dan $P(d)$ adalah probabilitas d. Untuk mencari probabilitas terbesar suatu musik didalam suatu kelas dapat dituliskan sebagai

$$C_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c) \quad (5)$$

C_{MAP} merupakan kelas dengan probabilitas d pada kelas c terbesar diantara semua kelas.

Untuk penelitian ini metode naive bayes yang diusulkan adalah multinominal naive bayes. Metode ini merupakan perkembangan dari metode naive bayes. Dengan metode ini kita tidak akan memperhatikan keterkaitan makna 2 kata atau lebih, tiap kata akan menjadi suatu ciri tersendiri. Metode ini dipilih karena dalam lagu daerah, setiap daerah umumnya mempunyai bahasa yang berbeda. Dengan memperhatikan sifat ini, metode ini dipilih untuk memperringan komputasi yang dilakukan.

3.5. Pengujian

Pengujian dilakukan dengan menggunakan data test yang telah dipisahkan sebelumnya sesudah preprocessing.

3.6. Evaluasi

3.6.1. F-Measure

F-Measure atau F-Score, mencari mean yang sesuai dengan precision dan recall. Dimana precision adalah jumlah persentase kebenaran pada kelas, dan recall adalah persentase kebenaran pada test data.

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \quad (6)$$

3.6.2. Confusion Matrix

Confusion matrix (Liu, Xiang, & Cai, 2009) (Patra, Das, & Bandyopadhyay, 2013). Confusion matrix merupakan matrix 2 dimensi yang berisi tiap kelas pada tiap dimensinya. Kolom biasanya merepresentasikan kelas asli, sedangkan baris merepresentasikan kelas yang diprediksikan.

4. EKSPERIMEN

Untuk mengetahui efektifitas dari metofe tf*idf dengan kombinasi metode klasifikasi naïve bayes kami melakukan serangkaian eksperimen. Lirik lagu yang kami gunakan sejumlah 90 lirik lagu daerah Indonesia. Untuk mendukung akurasi percobaan kami menggunakan rerata dari 10 kali cross validation yang dilakukan.

Tf*idf digunakan untuk mendapatkan fitur ciri dikarenakan faktor bahasa daerah di Indonesia yang masih belum tersedianya kamus kata bahasa daerah diindonesia. Tf*idf akan secara otomatis melakukan pembobotan terhadap kata yang terdapat pada lirik sehingga bisa dijadikan ciri untuk klasifikasi.

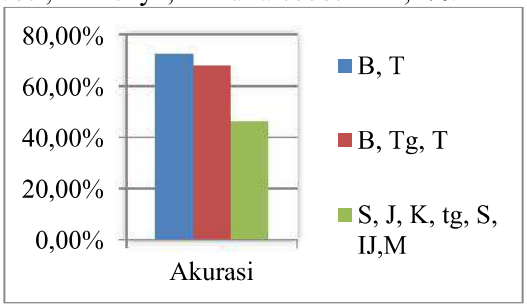
Naïve bayes digunakan untuk melakukan klasifikasi dikarenakan metode ini merupakan metode yang sudah dipercaya untuk melakukan klasifikasi lirik, dan membuahkan hasil yang memuaskan pada klasifikasi lagu indonesia. Kami menggunakan WEKA (Hall, et al., 2009) sebagai tools klasifikasi yang memang sudah banyak digunakan untuk klasifikasi pada bidang MIR.

Pada penelitian ini kami mengklasifikasikan lagu daerah berdasarkan daerah asal lagu. Set eksperimen yang kami gunakan adalah memisahkan berdasar daerah:

1. Indonesia Barat(Jawa, Bali, Sumatra, Kalimantan), Indonesia Timur(Irian jaya, Maluku, Sulawesi, NTT&NTB) (BT);
2. Indonesia Barat(Sumatra, Jawa, Kalimantan), Indonesia Tengah(Sulawesi, Bali, NTT&NTB), Indonesia Timur(Irian Jaya, Maluku) (BTgT); Sumatra, Jawa, Kalimantan, Bali, indonesia tengah(Maluku, NTT, NTB), Sulawesi, IrianJaya,(SJKBTgSI).

5. HASIL PENELITIAN

Berdasarkan set daerah yang ada, yang memiliki tingkat akurasi paling tinggi adalah set daerah indonesia Timur dan barat dengan F-Score sebesar 73,4%. Sedangkan set daerah dengan akurasi paling rendah adalah set daerah Sumatra, Jawa, Kalimantan, indonesia tengah(Bali, NTT&NTB), Sulawesi, IrianJaya, Maluku sebesar 46,6%.



Gambar 13. Grafik tingkat akurasi pada tiap set.

Dapat dilihat pada gambar 2 bahwa semakin banyak kelass yang dilibatkan, maka tingkat akurasi akan semakin menurun. Hal ini kemungkinan disebabkan oleh tidak meratanya pesebaran data latih dan data uji pada tiap kelas.

	S	J	K	B	Tg	S	I
S	**22**	3	1				
J	6	**14**	1	1	2	1	1
K	3	4	**5**			1	1
B	2	2	1	**0**			3
Tg	1	3		1	**0**	1	
S	3	1				**3**	2
I	1					3	

Tabel 1. Confusion Matrix SJKBTgSI

Data yang tersedia adalah 90 data dengan detail 26 lagu sumatra, 28 lagu jawa, 14 lagu kalimantan, 8 lagu bali, 9 lagu sulawesi, 4 lagu irian, 3 lagu nusa tenggara barat dan timur,dan 3 lagu maluku. Pesebaran data yang tidak seimbang ini bisa menjadi penyebab, karena model akan dihasilkan lebih condong ke arah kelas yang memiliki data lebih banyak.

	Barat	Timur
Barat	**52**	16
Timur	10	**17**

Tabel 2. Confusion matrix untuk set data BT

Berdasarkan Gambar 3, Dapat dilihat bahwa pengklasifikasian pada lagu indonesia timur hampir 50% dikategorikan lagu indonesia barat. Hal ini dimungkinkan karena jumlah lagu untuk yang digunakan pada set Indonesia Timur lebih sedikit dari pada indonesia Barat, sehingga ragam kata pada kelas indonesia Barat lebih beragam dan menjadikan peluang lagu teridentifikasi pada kelas indonesia timur berkurang.

Faktor bahasa yang jauh berbeda juga mempengaruhi akurasi. Lirik lagu daerah yang kami dapatkan sangat jauh berbeda antar lagu yang satu dengan yang lain. Jumlah ragam kata yang terlalu banyak mengakibatkan data kata yang ada pada data test tidak dapat dicari kesamaannya pada model training sehingga tingkat akurasi pun menjadi rendah.

6. KESIMPULAN DAN SARAN

1. Kesimpulan dari penelitian ini adalah metode tf*idf dan naive bayes dapat mengklasifikasikan lagu daerah dengan akurasi sangat tinggi yaitu 72,63%.
2. Eksperimen yang dilakukan disini hanya melihat dari sisi aspek linguistik pada lagu. Untuk meningkatkan hasil akurasi, fitur lain yang ada pada audio lagu bisa dijadikan pertimbangan.

Untuk itu diperlukan data audio lagu daerah sebagai tambahan pada lirik.

3. Jumlah kata yang ada pada lirik pun beragam ada yang 50 kata, 80 kata, dan sebagainya.
4. Metode klasifikasi yang digunakan pada penelitian ini adalah dengan menggunakan metode pembelajaran terbimbing naive bayes yang berdasarkan probabilitas. Akan tetapi karena lirik lagu daerah yang didapat amat sangat berbeda (terutama pada daerah indonesia timur), maka akan membuat kemungkinan besar untuk menghasilkan 0 pada probabilitas dan menjadikan pengklasifikasian akan menjadi kurang akurat.

References

Cyril Laurier, J. G. (2008). Multimodal Mood Classification using Audio and Lyris. *In Proceeding of the 7th International Conference on Machine Learning and Application (ICMLA' 08)*, 1-6.

Dewi, L. M. (2011). Incorporating Global Information In Follk Song Classification System. 1-53.

Frans Wiering, R. C. (2009). Modelling Folksong Melodies. *ITERDISCIPLINARY SCIENCE REVIEWS*, 1-18.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1).

Liu, Y., Xiang, Q., & Cai, L. (2009). Cultural Style Based Music Classification of Audio Signals. *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference* (pp. 57-60). Taipei: IEEE.

Matt McVicar, T. F. (2011). MINING THE CORRELATION BETWEEN LYRICAL AND AUDIO FEATURES AND THE EMERGENCE OF MOOD. *12th International Society for Music Information Retrieval* , 1-6.

Min-Yen Kan, Y. W. (2008). LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 16*, 338.

Patra, B. G., Das, D., & Bandyopadhyay, S. (2013). Automatic Music Mood Classification of Hindi Songs. *Proceedings of the 3rd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2013), IJCNLP 2013*, (pp. 24-28). Nagoya,Japan.

Rauber, R. m. (2011). MUSICAL GENRE CLASSIFICATION BY ENSEMBLES OF AUDIO AND LYRICS FEATURES. *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, 675.

Robert Neumayer, A. R. (2007). Multi-modal music information retrieval - visualisation and evaluation of clusterings by both audio and lyrics. *In Proceedings of the 8th Conference Recherche d'Information Assistée par Ordinateur (RIAO'07)*, 1-20.

S.Daudu. (2009). Problem and Prospect of Folk Media Usage for Agricultural Extension Service Delivery in Benue State, Nigeria. 1-6.

Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation of Musical Features For Emotion Classification. *13th International Society for Music Information Retrival Confrence (ISMIR2012).*

Xiaohu, J. D. (2009). LYRIC TEXT MINING IN MUSIC MOOD CLASSIFICATION. *10th International Society for Music Iformation Retrieval Conference (ISMIR 2009)*, 1-6.

Xing Wang, X. C. (2011). MUSIC EMOTION CLASSIFICATION OF CHINESE SONG BASED ON LYRICS USING TF*IDF AND RHYME. *12th International Society for Music Information Retrieval Conference(ISMIR 2011)*, 1-6.

Zaanen, M. v., & Kanters, P. (2010). Automatic Mood Classification Using TF*IDF Based on Lyrics. *Proceedings of the 11th International Society for Music Information Retrieval Conference*, (pp. 75-80). Utrecht, The Netherlands.