Manuscript Submitted	22/5/2022
Accepted	20/6/2022
Published	29/6/2022

Menangani Ketaksaan dalam Transliterasi Mesin Jawi - Rumi menggunakan Pengelasan Naive Bayes Multinomial (NBM)

Che Wan Shamsul Bahri Che Wan Ahmad
Fakulti Sains dan Teknologi Maklumat
Kolej Universiti Islam Antarabangsa Selangor (KUIS)
Bandar Seri Putra, Bangi, Selangor, Malaysia
cwshamsul@kuis.edu.my

Khairuddin Omar, Mohammad Faidzul Nasruddin & Mohd Zamri Murah
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia(UKM)
{ko,mfn, zamri}@ukm.edu.my

Abstract

This paper discusses the problem of ambiguity in Jawi - Rumi machine transliteration for Jawi homograph words. Machine transliteration (MT) is the process of converting a script from source text to target text automatically. In the context of Malay MT for Jawi - Rumi, there are difficulties in obtaining high -accuracy transliteration of homographical Jawi words. Homographs are words that are the same spelling, but have different meanings and pronunciations. In the old Jawi spelling there were many homograph words, while it was successfully reduced when "Pedoman Ejaan Jawi yang Disempurnakan" (PEJYD) was first introduced by Dewan Bahasa dan Pustaka (DBP) in 1986. The main issue in the study of Malay Jawi - Rumi machine transliteration was word inaccuracy when the Jawi word is transliterated to Rumi. For example, the word "بيرو" can be transliterated to 'biru'(blue) or 'biro'(bureau), the word "بيليق" can be transliterated to 'bilik'(room) or 'belek'(turn around). This paper proposes that the Multinomial Naive Bayes (NBM) classification method be used for homograph unambiguity for TM Jawi - Rumi. Test results found that the accuracy of using this method can reach up to 67 percent.

***Keywords**: homograph, natural language processing (NLP), Jawi, machine transliteration*

Abstrak

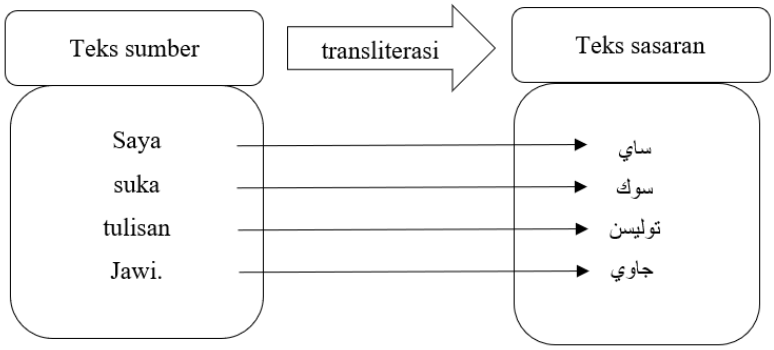
Kertas ini membincangkan masalah ketaksaan dalam transliterasi mesin Jawi – Rumi bagi perkataan homograf Jawi. Transliterasi mesin (TM) adalah proses menukar skrip daripada teks sumber kepada teks sasaran secara automatik. Dalam konteks TM Bahasa Melayu (BM) Jawi - Rumi, terdapat kesukaran untuk mendapatkan transliterasi yang berketepatan tinggi bagi perkataan Jawi yang homograf. Homograf adalah perkataan yang sama ejaannya, tetapi mempunyai makna dan sebutan berbeza. Dalam ejaan Jawi lama terdapat banyak perkataan homograf, manakala ia berjaya dikurang apabila Pedoman Ejaan Jawi yang Disempurnakan (PEJYD) mula diperkenalkan oleh Dewan Bahasa dan Pustaka (DBP) pada tahun 1986. Isu utama dalam kajian transliterasi mesin BM Jawi - Rumi adalah ketaksaan perkataan apabila perkataan Jawi ditransliterasi kepada Rumi. Contohnya perkataan "بيرو" boleh ditransliterasi kepada biru atau biro, perkataan "بيليق" boleh ditransliterasi kepada bilik atau belek. Kertas ini mencadangkan kaedah pengkelasan Naive Bayes Multinomial (NBM) digunakan untuk penyahtaksaan homograf bagi TM Jawi - Rumi. Hasil ujian mendapati ketepatan menggunakan kaedah ini boleh mencapai sehingga 67 peratus.

.

Kata kunci: homograf, pemprosesan bahasa tabii (PBT), Jawi, transliterasi mesin

1. Pengenalan

Menurut Kamus Dewan Edisi Keempat (2010), transliterasi adalah penukaran huruf atau perkataan dan sebagainya daripada abjad sesuatu tulisan (contohnya tulisan Arab) kepada huruf yang selaras bunyinya dan sebagainya dalam abjad sistem tulisan lain (contohnya Rumi). Dengan erti kata lain, transliterasi adalah satu proses menyalin semula perkataan yang ditulis daripada satu bahasa (bahasa sumber) kepada bahasa lain (bahasa sasaran) dengan memelihara artikulasinya sebagaimana dalam bahasa asal (Malik et al., 2009). Menurut Virga & Khudanpur (2003), artikulasi adalah pembentukan suara yang jelas dan terang. Rajah 1 menunjukkan proses asas transliterasi dalam Bahasa Melayu bagi pertukaran teks daripada aksara Rumi kepada aksara Jawi.



Rajah 1. Transliterasi Jawi Rumi
Sumber: Shamsul et al. (2012)

2. Isu-isu dalam Transliterasi Mesin

TM merupakan salah satu sub-bidang pengkomputeran linguistik dan keperluan dalam pemprosesan bahasa untuk tugas bagi sesuatu sifat bahasa tertentu. Walaupun banyak kajian memperkenalkan kaedah statistik sebagai penyelesaian kegunaan umum untuk kedua-dua sub-bidang terjemahan dan transliterasi, pengetahuan khusus dalam bahasa tersebut juga adalah satu keperluan. Dalam konteks TM Jawi- Rumi, pengetahuan dalam kaedah penulisan sistem ejaan Jawi adalah sebagai pra-syarat yang perlu ada (Yonhendri, 2008).

Permasalahan utama dalam kajian TM adalah untuk mendapatkan hasil transliterasi yang berketepatan tinggi. Pendekatan transliterasi sebelum ini boleh dibahagikan kepada tiga kategori utama: pendekatan berasaskan grafem; pendekatan berasaskan fonetik; dan pendekatan yang menggabungkan ciri kedua-duanya(Karimi et al., 2006). Namun terdapat juga pelbagai model, teknik dan kaedah yang digunakan untuk mencapai tujuan tersebut seperti transliterasi berasaskan petua, pembelajaran mesin (PM), pendekatan berstatistik dan pendekatan perlombongan Web (Zhou et al., 2008).

Dalam TM bahasa asing, skrip yang berbeza antara bahasa sumber dan bahasa sasaran merupakan halangan pertama yang perlu diatasi oleh sistem transliterasi. Skrip ialah perwakilan satu atau lebih sistem tulisan dan terdiri daripada simbol yang digunakan untuk mewakili teks. Satu skrip boleh digunakan untuk beberapa bahasa yang berbeza; sebagai contoh, skrip Latin merangkumi seluruh Eropah Barat, dan skrip Arab digunakan untuk bahasa Arab, dan beberapa bahasa bukan Semitik yang ditulis dalam abjad Arab termasuk Parsi, Urdu, Pashto, Melayu (Jawi) dan Balti. Sebaliknya, sesetengah bahasa bertulis memerlukan berbilang skrip (Karimi 2008; Karimi et al. 2006, 2011). Contohnya, bahasa Jepun ditulis dalam sekurang-kurangnya tiga skrip: suku kata Hiragana dan Katakana dan ideograf Kanji. Pemprosesan pengkomputeran skrip bahasa yang berbeza ini memerlukan keupayaan untuk mengendalikan pengekodan aksara yang berbeza.

Walaupun sesetengah skrip ditulis menggunakan aksara yang berasingan (seperti Latin), yang lain memperkenalkan bentuk perantaraan untuk aksara yang terdapat di tengah-tengah perkataan. Sebagai contoh, dalam skrip Parsi beberapa huruf menukar bentuknya berdasarkan kedudukannya dalam perkataan, "پ" [p] jika tidak bersambung, dan "پـ" [p] pada awal perkataan yang bersambung.

Satu lagi aspek penting dalam skrip bahasa ialah arah penulisannya. Sesetengah bahasa ditulis dari kanan-ke-kiri (RTL) dan beberapa bahasa ditulis dari kiri-ke-kanan (LTR). Contohnya, skrip Parsi, Arab, Ibrani dan Taana ialah RTL, manakala skrip bahasa Inggeris dan bahasa lain yang menggunakan abjad Latin ialah LTR. Secara umum, TM mampu memanipulasi aksara perkataan dan harus direka bentuk dengan teliti untuk memproses skrip bahasa sumber dan bahasa sasaran dengan mengambil kira semua spesifikasi yang dinyatakan di atas.

Bunyi yang hilang adalah antara cabaran dalam TM. Pelbagai bahasa di dunia ini mempunyai struktur bunyi yang tersendiri dan simbol skrip bahasa sepadan dengan bunyi tersebut. Jika terdapat bunyi yang hilang dalam huruf sesuatu bahasa, bunyi tunggal diwakili menggunakan digraf dan trigraf. Sebagai contoh, digraf Inggeris "sh" sepadan dengan bunyi [S]. Terjemahan bunyi merentas bahasa dengan transliterasi akan menghasilkan bunyi baharu kepada bahasa sasaran, yang tidak semestinya dimuatkan oleh bahasa sasaran

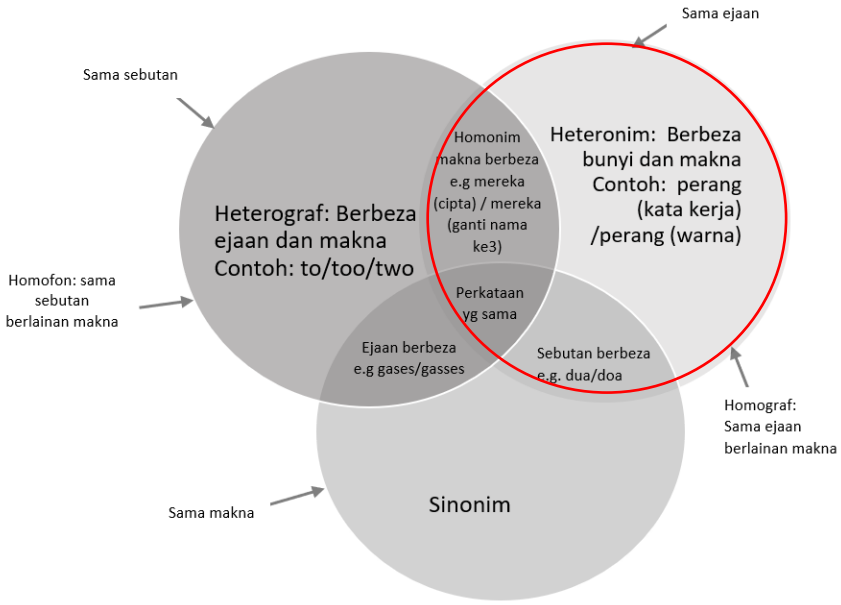
Berdasarkan kajian-kajian lepas dalam penyelidikan Pemprosesan Bahasa Tabii (PBT) yang melibatkan transliterasi Bahasa Melayu Jawi - Rumi, antara isu-isu yang dikenalpasti adalah seperti berikut :

Rajah 2. Permasalahan dalam TM Jawi - Rumi

Isu yang cuba diketengahkan dalam kertas ini adalah berkaitan isu ketaksaan transliterasi perkataan Jawi kepada Rumi.

3. Perkataan ambiguiti dalam Jawi

Homograf adalah perkataan mempunyai ejaan sama tetapi sebutan dan maksud yang berbeza (Adi Yasran & Hashim, 2008) sebagaimana ditunjukkan pada Rajah 3 di bawah. Oleh itu, hasil transliterasi Jawi kepada ejaan Rumi juga akan menghasilkan lebih daripada satu padanan perkataan.



Rajah 3 Perkataan dengan ejaan, sebutan dan makna yang berbeza

Ia berbeza dengan Jawi moden yang kebanyakannya mempunyai persamaan dengan ejaan Rumi dan mengandungi peraturan yang kemas dan sistematik (Shamsul Bahri et al., 2012; Hamdan, 1999, 2013). Sebagai contoh perkataan yang homograf dalam ejaan Jawi lama dan tidak berlaku homograf dalam Jawi baru sebagaimana Jadual 1 di bawah.

Jadual 1. Perkataan homograf hakiki dalam Jawi moden

Jawi lama	*Jawi baru*	*Rumi*
كمڤوڠ	كامڤوڠ	kampung
	كمڤوڠ	kempung
چڠكوڠ	چاڠكوڠ	cangkung
	چڠكوڠ	cengkung
كنچيڠ	كانچيڠ	kancing
	كنچيڠ	kencing
لبه	لبيه	lebih
	لبه	lebah
كام	كامو	kamu
	كامي	kami

Namun begitu, masih terdapat perkataan homograf dalam ejaan Jawi baru seperti contoh pada Jadual 2 berikut:

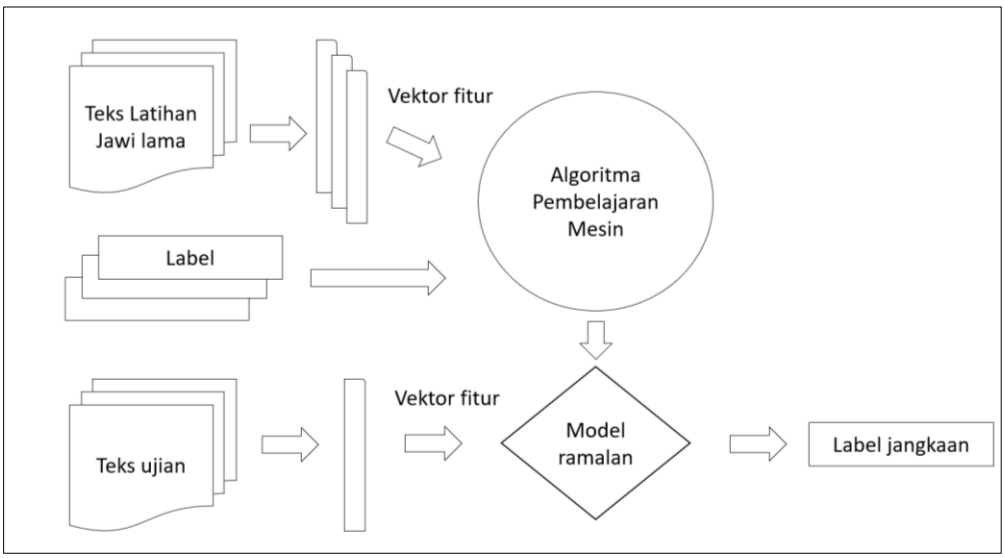
Jadual 2. Perkataan homograf hakiki dalam Jawi moden

Jawi	***Rumi***	
ڬولوڠ	gulung	golong
بيليق	bilik	belek
كليه	kuliah	kelih
بيرو	biru	biro
سمبيلن	sambilan	sembilan
مركه	merekah	markah
سيسي	sisi	sesi

4. Cadangan penyelesaian

Di dalam penyelidikan ini, *textblob* dan *Python* digunakan untuk pengelasan NBM. Pengelasan teks adalah sistem yang mengklasifikasikan sesuatu teks dan membahagikannya ke dalam kategori yang berbeza sebagaimana ditunjukkan pada Rajah 4 di bawah.

Vektor fitur yang mewakili teks mengandungi kebarangkalian kemunculan teks dalam kategori perkataan tertentu supaya algoritma boleh mengira kemungkinan teks tersebut tergolong dalam kategori tersebut. Sebaik sahaja kemungkinan pelbagai kelas diperoleh, kelas yang mempunyai kemungkinan maksimum akan dipilih bagi mengkategorikan teks tersebut.



Rajah 4 Rangka kerja pengelasan teks Jawi menggunakan NBM

Metod *textblob* iaitu salah satu metod yang terdapat dalam perpustakaan *Python* (2 dan 3) untuk pemprosesan data teks dengan cara yang mudah. Ia menyediakan *'Application Programming Interface'* (API) yang konsisten untuk melakukan tugas PBT seperti penandaan bahagian ucapan, pengekstrakan frasa nama, analisis sentimen, pengklasifikasian dan terjemahan.

Beberapa set data latihan dan ujian dibangunkan terlebih dahulu. Rajah 5 di bawah menunjukkan data latihan dan data ujian untuk perkataan ‘بيرو' (biru/biro) yang telah disediakan dengan kedua-dua data tersebut perlu dilabelkan mengikut kelas masing-masing.

```
latihan = [
('لاڠيت دان لاوت ورنا بيرو', 'biru'),
('باجو ساي بيرو', 'biru'),
('سموا لجنة واجب مماكاي باجو سراكم ورنا بيرو', 'biru'),
('كتيك هاري ترڠ لاڠيت بيرو', 'biru'),
('لاڠيت برورنا بيرو كتيك هاري چره', 'biru'),
('ورنا بيرو ڤنتيڠ باكي مناڠني مسئله تكنن داره تيڠكي يڠ ممبريكن كسن لبيه تنڠ', 'biru'),
('اديق مروڤاكن ساله سأورڠ اتليت اهلي رومه سوكن بيرو', 'biru'),
('ورنا لاوت بيرو', 'biru'),
('اهلي بيرو ايت ممندڠ ك لاڠيت كتيك ڤرباريسن داداكن', 'biro'),
('اهلي بيرو دان اهلي لجنة', 'biro'),
('سموا بيرو دواجبكن مماكاي باجو سراكم ورنا بيرو', 'biro'),
('بيرو سوكن دان ريكرياسي برتيكد ممبري كفهمن دان كسدران تنتڠ كڤنتيڠن سوكن', 'biro'),
("بيرو تاجان امت ڤنتيڠ دالم سسبواه ڤرساتوان", 'biro'),
('ستياڤ بيرو دان لجنة ممڤوڤاءي توكس يڠ ڤنتيڠ', 'biro'),
]
ujian = [
('باجو اديق ورنا بيرو', 'biru'),
('اهلي ڤرساتوان يڠ ممكڠ جاوتن سباكاي بيرو مندافت باڤق كلبيهن', 'biro'),
('ساي ممكڠ جاوتن بيرو تاجان دالم ڤرساتوان', 'biro'),
("لاڠيت دان لاوت برورنا بيرو", 'biru'),
('ڤين چنكو ورنا بيرو تله هيلڠ', 'biru'),
('ستياڤ بيرو ڤرلو تكس دالم ممبري ارهن', 'biro'),
("بيرو دان لجنة ملقساناكن فوغسي سڤرتي سبواه جاوتنكواس", 'biro'),
('بيرو ترجمهن نكارا ڤرلو دتوبوهكن سكرا', 'biro'),
("بيرو سوكن امت ڤنتيڠ", 'biro'),
("بيرو دان لجنة دسارنكن منجالنكن توكس دڠن تلوس", 'biro')
]]
```

Rajah 5 Contoh data latihan dan data ujian untuk perkataan ‘بيرو’

Rajah 6 di bawah adalah contoh panggilan metod *classify()* untuk mengelaskan teks dengan menggunakan pengelas tersebut.

```
classify("باجو اديق ورنا بيرو")
classify("اهلي ڤرساتوان يڠ ممكڠ جاوتن سباكاي بيرو مندافت باڤق كلبيهن")
classify("بيرو سوكن امت ڤنتيڠ")
```

Rajah 6 Metod *classify()*

Untuk melakukan penyahtaksaaan homograf Jawi, satu fungsi yang diberi nama fungsi *Nyah_kabur* dibangunkan dengan menggunakan beberapa metod yang terdapat di dalam Pustaka *Python*. Jadual 3 di bawah adalah contoh data latihan untuk penyahtaksaan homograf bagi perkataan “بيرو”. Sekiranya data latihan dapat disediakan dengan jumlah yang banyak, maka tetingkap konteks menghasilkan lebih banyak kata kunci untuk mendapat ramalan yang lebih tepat.

Dalam contoh di bawah, perkataan “بيرو” mempunyai dua kebarangkalian transliterasi ke Rumi sama ada (A) biro atau (B) biru. Dalam contoh di bawah, sebanyak 12 contoh ayat latihan diambil daripada Korpus Dewan Bahasa dan Pustaka (2021). Konteks diambil dari kata kunci di sekitarnya; dalam kes ini, tetingkap konteks yang diset saiznya kepada 3, iaitu P[0], P[1], dan P[2] diambil dari kata kunci yang mendahului perkataan homograf dalam urutan yang muncul dalam teks, dan P[3], P[4], dan P[5] diambil dari kata kunci yang melepasi perkataan homograf dalam teks.

Jadual 3 Contoh data latihan bagi berkataan homograf "بيرو"

Kelas				P[5]	P[4]	P[3]	homo	P[2]	P[1]	P[0]	
B						مودا	بيرو	باجو	مماكاي	بنتوان	قوليس
B						كَلڤ	بيرو	سلوار	دان		
B			چره	هاري	كتيك	ترڠ	بيرو	برورنا	لاڠيت		
A			سمولا	دكاجي	ڬوامن	بنتوان	بيرو	فوڠسي	سوڤاي	منچادڠكن	
A	داداكن	قرباريسن	كتيك	كلاڠيت	ممندڠ	ايت	بيرو	اهلي			
B						كَلڤ	بيرو	برورنا	رودا	كاسوت	سڤاسڠ
B					اون	تنڤا	بيرو	لاڠيت			
A	كفهامن	ممبري	برتيكد	ريكرياسي	دان	سوكن	بيرو	ڤڠروسي			
A					اوسهاون	ڤمباڠونن	بيرو	ڤڠروسي	جوڬ	عبدالغاني	محمد
B			كچيل	بوڠا	دان	لمبوت	بيرو	برورنا	تيدور	باجو	
A	ڤولڠ	تيدق	ماسيه	تتاڤي	ڬوامن	بنتوان	بيرو	درڤد	بنتوان	مندافتكن	
A	ڤرساتوان	سسبواه	دالم	ڤنتيڠ	امت	تاجأن	بيرو	فوڠسي			

Menanda atau melabelkan set data latihan agar berada dalam kelas yang betul adalah penting dan perlu disemak dan disahkan oleh pakar Jawi. Berikut adalah entri leksikon penuh untuk perkataan "بيرو" seperti yang dihasilkan oleh pengelas NBM sebagaimana contoh data dalam Jadual 3 di atas.

Perkataan yang termasuk dalam kelas A (biro) adalah seperti berikut –

دكاجي، ڬوامن، بنتوان، فوڠسي، سوڤاي، منچادڠكن، كلاڠيت، ممندڠ، ايت، اهلي، ريكرياسي، دان، سوكن، ڤڠروسي، اوسهاون، ڤمباڠونن، ڤڠروسي، جوڬ، عبدالغاني، تتاڤي، ڬوامن، بنتوان، درڤد، بنتوان، مندافتكن، ڤنتيڠ، امت، تاجأن، فوڠسي

Perkataan yang termasuk dalam kelas B (biru) pula adalah seperti berikut–

مودا، باجو، مماكاي، بنتوان، كَلڤ، سلوار، دان، هاري، كتيك، ترڠ، برورنا، لاڠيت، كَلڤ، برورنا، رودا، كاسوتاون، تنڤا، لاڠيت، بوڠا، دان، لمبوت، برورنا، تيدور، باجو

Senarai perkataan tersebut dibahagikan kepada dua kelas sebagaimana Jadual 4 berikut dengan mengambil kira kekerapan perkataan tersebut dan jumlah keseluruhan perkataan bagi setiap kelas.

Jadual 4 Senarai perkataan bagi setiap kelas

Kelas A	
perkataan	***kekerapan***
امت	1
اهلي	1
اوسهاون	1
ايت	1
بنتوان	3
تاجأن	1
تتاڤي	1
جوڬ	1
دان	1
درڤد	1
دكاجي	1
ريكرياسي	1
سوڤاي	1
سوكن	1
عبدالغاني	1
فوغسي	2
ڤغروسي	2
ڤمباغونن	1
ڤنتيغ	1
كلاغيت	1
ڬوامن	2
ممندغ	1
منچادغكن	1
مندافتكن	1
Jumlah	29

Kelas B	
perkataan	***kekerapan***
باجو	2
برورنا	3
بنتوان	1
بوغا	1
ترغ	1
تنڤا	1
تيدور	1
دان	2
رودا	1
سلوار	1
كتيك	1
كاسوتاون	1
ڬلڤ	2
لاغيت	2
لمبوت	1
مماكاي	1
مودا	1
هاري	1
Jumlah	24

Katakan teks untuk ujian sebagai contoh di bawah

" ڤماكاين اونيفورم كميجا كونيغ أير دان سلوار بيرو ڬلڤ دغن تندا نام"
(pemakaian uniform kemeja kuning air dan seluar biru gelap dengan tanda nama)

Dalam formula NBM, kebanyakan kebarangkalian bersyarat didarabkan. Untuk mengelakkan berlaku nombor perpuluhan semakin kecil dalam sistem komputer, pelaksanaan pengiraan dalam ruang log digunakan dengan menjumlahkan logaritma kebarangkalian dan bukannya mendarab kebarangkalian (Manning et al., 2008). Fungsi logaritma adalah monotonik dengan $\log(x \times y) = \log(x) + \log(y)$.

Hasilnya setelah memeriksa set data yang disediakan oleh struktur penyahtaksaan dalam entri leksikon dengan mengira kebarangkalian log adalah seperti Jadual 5 berikut:

Jadual 5 Pengiraan kebarangkalian log setiap perkataan

Perkataan	*Kelas A - biro*		*Kelas B - biru*	
ڤماکاين	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
اونيفورم	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
كميجا	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
كونيڠ	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
أير	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
دان	log((1/29)+1)	≈ 0.0147	log((2/24)+1)	≈ 0.0348
سلوار	log((0/29)+1)	≈ 0	log((1/24)+1)	≈ 0.0177
بيرو	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
ڬلڤ	log((0/29)+1)	≈ 0	log((2/24)+1)	≈ 0.0348
دڠن	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
تندا	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
نام	log((0/29)+1)	≈ 0	log((0/24)+1)	≈ 0
	Jumlah	**0.0147**		**0.0873**

Dengan membandingkan jumlah skor pada Jadual 5 di atas, kita dapati bahawa nilai skor kelas B adalah lebih besar daripada skor kelas A iaitu 0.0873 > 0.0147. Oleh itu, fungsi ramalan iaitu fungsi *Nyah_kabur* akan memilih kelas B berdasarkan contoh ujian yang dilakukan. Justifikasi keputusan ini adalah bahawa kejadian دان, سلوار, dan ڬلڤ dalam data latihan kelas B mengatasi kejadian دان dalam data latihan kelas A.

5. Pengujian

Terdapat isu utama dalam TM Jawi Rumi adalah bagaimana ramalan perlu dilakukan untuk memilih padanan yang tepat bagi perkataan Rumi yang mempunyai lebih daripada satu padanan bagi satu perkataan Jawi sebagaimana dibincangkan dalam perkara 5 di atas.

Eskperimen yang dilakukan untuk menguji ketepatan kaedah yang digunakan untuk perkataan homograf tersebut. Perkataan homograf yang dipilih adalah berdasarkan perkataan homograf hakiki. Homograf jenis ini masih berlaku dalam ejaan Jawi baru. Sebanyak 12 perkataan homograf yang dipilih dalam kajian ini adalah seperti pada Jadual 6 di bawah:

Jadual 6 Perkataan homograf Jawi

Bil	*Perkataan homograf hakiki*
1	بيرو (biru – biro)
2	ڬولوڠ (gulung – golong)
3	بوروڠ (burung – borong)
4	امبين (amben – ambin)
5	تونتون (tonton – tuntun)
6	سوروت (sorot – surut)
7	سيسي (sesi – sisi)
8	سمبيلن (sembilan – sambilan)
9	مركه (merekah – markah)
10	كوكو (kuku – koko)
11	كليه (kelih – kuliah)
12	اسا (esa – asa)

Rajah 7 di bawah adalah algoritma ringkas fungsi *Nyah_kabur* untuk penyahtaksaan homograf Jawi.

Algoritma untuk fungsi Nyah_kabur untuk penyahtaksaan homograf Jawi

```
Keperluan: Data latihan Jawi J1, Data ujian Jawi J2

#Memuatkan Data dan Membuat Pengelas
Keperluan: Data latihan Jawi J1, Data ujian Jawi J2

#Mengelaskan Teks
Panggil metod classify (teks) untuk menggunakan pengelas.
Dapatkan nilai taburan kebarangkalian kelas

#Mengelaskan TeksBlob
Memasukkan pengelas ke dalam konstruktor TextBlob
Panggil metod classify()

#Menilai Pengelas
Panggil metod accuracy (data ujian, J2).

Paparkan senarai fitur yang paling berinformatif
```

Rajah 7 Algoritma fungsi penyahtaksaan homograf Jawi

Selepas set data latihan dan data ujian dibangunkan, barulah kita boleh membuat pengujian untuk perkataan homograf tersebut. Bagi perkataan "بيرو", nilai taburan kebarangkalian kelas diperolehi bagi setiap ayat yang diuji adalah seperti Jadual 7 berikut:

Jadual 7 Nilai taburan kebarangkalian kelas untuk perkataan "بيرو"

Ayat yang mengandungi perkataan Jawi homograf	***Taburan kebarangkalian kelas***
اهلي ڤرساتوان يڠ ممڬڠ جاوتن سباڬاي اهلي بيرو مندافت باڽق كلبيهن (Ahli persatuan yang memegang jawatan sebagai ahli **biro** mendapat banyak kelebihan)	biro: 0.61 biru: 0.39
اهلي بيرو سوكن ڤنتيڠ د مسجد قريه باتو 10 چراس (Ahli **biro** sukan penting di Masjid Kariah Batu 10 Cheras)	biro: 0.55 biru: 0.45
باجو اديق ورنا بيرو (Baju adik warna **biru**)	biro: 0.02 biru: 0.98

6. Perbincangan

Didapati permasalahan ketaksaan transliterasi perkataan Jawi ke Rumi dapat diselesaikan dengan menggunakan pendekatan NBM. Eksperimen yang dijalankan terhadap 12 perkataan Jawi homograf mendapati nilai ketepatan boleh mencapai purata sehingga 67 peratus.

Jadual 8 Nilai ketepatan data latihan bagi setiap perkataan homograf Jawi

Perkataan homograf	*Nilai ketepatan data latihan*
biru – biro (بيرو)	0.9
gulung – golong (ڬولوڠ)	0.6
burung – burong (بوروڠ)	0.64
amben – ambin (امبين)	0.67
tonton – tuntun (تونتون)	0.6
sorot – surut (سوروت)	0.5
sesi – sisi (سيسي)	1.0
sembilan – sambilan (سمبيلن)	0.6
merekah – markah (مركه)	0.6
kuku – koko (كوكو)	0.6
kelih – kuliah (كليه)	0.5
Esa – asa (اسا)	0.83
Purata	**0.67**

7. Kesimpulan

Kertas ini berjaya membuktikan bahawa dengan menggunakan pendekatan NBM, penyahtaksaan perkataan homograf dalam TM dapat dilakukan dengan baik. Namun, cadangan untuk kajian pada masa depan, kaedah-kaedah lain pengelasan teks dalam pembelajaran mesin seperti *Support Vector Machines* (*SVM)* dan *Deep Learning* seperti *Convolutional Neural Networks (CNN)* dan *Recurrent Neural Networks (RNN*) yang mana juga boleh dijalankan untuk menguji keberkesanannya bagi transliterasi mesin Jawi kepada Rumi.

Rujukan

Adi Yasran, A. A., & Hashim, H. M. (2008). Isu Homograf dan Cabarannya dalam Usaha Pelestarian Tulisan Jawi. *Jurnal ASWARA (Akademi Seni Budaya dan Warisan Kebangsaan)*, *3*(1), 109–126.

Che Wan Shamsul Bahri, C. A., Khairuddin, O., Mohammad Faidzul, N., Mohd Zamri, M., & Abd Rahman, K. (2012). Comparative Study Between Old and Modern Jawi Spelling: Case Study on Kitab Hidayah al-Salikin. *Proceeding of the 8th World Conference on Muslim Education, WorldCOME 2012*, hlm. 1-14.

Che Wan Shamsul Bahri, C. W. A., Khairuddin, O., Nasrudin, M. F., Mohd Zamri, M. M., & Sanusi, M. A. (2012). Isu-isu dalam transliterasi mesin manuskrip Melayu ejaan Jawi lama kepada Jawi baru. *Seminar Penyelidikan Jawi dan Manuskrip Melayu*, hlm. 169-179.

Dewan Bahasa dan Pustaka. (2021). *Korpus Dewan Bahasa dan Pustaka*. http://sbmb.dbp.gov.my/korpusdbp/Researchers/Search2.aspx

Hamdan, A. R. (1999). *Panduan Menulis dan Mengeja Jawi*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Hamdan, A. R. (2013). Aksara Jawi dari zaman kuno hingga zaman penjajahan. *Seminar perkaedahan Jawi: Evolusi tulisan Jawi.*, hlm. 1-13.

Kamus Dewan edisi 4. (2010). Kamus Dewan. *Kuala Lumpur: Dewan Bahasa dan Pustaka*.

Karimi, S., Turpin, A., & Scholer, F. (2006). English to Persian transliteration. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4209 LNCS*, 255–266. https://doi.org/10.1007/11880561_21

Malik, A., Besacier, L., Boitet, C., & Bhattacharyya, P. (2009). A hybrid model for Urdu Hindi transliteration. *Proceedings ofthe 2009 Named Entities Workshop, August*, hlm. 177-185. https://doi.org/10.3115/1699705.1699746 [8 Mei 2018].

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. *Cambridge: Cambridge University Press*.

Virga, P., & Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Language Applications. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, *SPEC. ISS.*, 365–366. https://doi.org/10.1145/860500.860503

Yonhendri. (2008). *Enjin Transliterasi Rumi Jawi*. Tesis Sarjana, Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia.

Zhou, Y., Huang, F., & Chen, H. (2008). Combining probability models and web mining models: a framework for proper name transliteration. *Information Technology and Management*, *9*(2), 91–103.