Pernormalan Teks Media Sosial Bahasa Melayu

Liew Kean Oon

Prof Dr. Shahrul Azman Mohd Noah

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Ciap Bahasa Melayu digunakan secara meluas oleh pengguna twitter kini terutamanya pengguna dalam negara Malaysia. Oleh itu, pernormalan teks ciap Bahasa Melayu amat diperlukan agar ciap Bahasa Melayu dapat diterjemahan ke dalam Bahasa Melayu yang piawai, supaya proses selanjutnya seperti analisis sentimen, pemproses teks dan analitik teks dapat dilaksanakan secara berkesan. Terdapat sesetengah penyelidik telah menjalankan proses pemprosesan bahasa tabii yang menumpukan kepada pernormalan teks media sosial ataupun ciap berbahasa Inggeris, sementara kajian penormalan ciap berbahasa Melayu masih terhad. Maka, projek ini mencadangkan kaedah berasakan kamus hibrid untuk menormalkan ciap Bahasa Melayu yang melibatkan singkataan perkataan ke perkataan asal.

1. PENGENALAN

Perkataan “kamus” berasal daripada bahasa Arab, iaitu qamus. Kamus asalnya merupakan sejenis buku rujukan yang menerangkan makna sesuatu kata dan berfungsi untuk membantu pengguna mengenal perkataan baharu. Selain menerangkan maksud perkataan, sesetengah kamus turut mengandungi panduan sebutan, etimologi dan contoh penggunaannya. Kini, penggunaan komputer dalam proses pengajaran dan pembelajaran telah meluas dengan pesat. Melalui teknologi internet, ianya telah membantu pelajar mengakses laman web ilmiah dan informasi pendidikan terkini secara mudah dan pantas. Nolan dan Martin (1994), menyatakan bahawa pembelajaran dalam suasana baru seperti menggunakan internet di sekolah telah membawa banyak perubahan kepada pelajar.

Mengenai pembelajaran yang sedia ada, pembelajaran kemahiran menulis menjadi satu cabaran kepada pelajar kerana kekurangan kemahiran dalam aspek pemilihan kata ataupun istilah. Kebanyakan pelajar kurang memahami penggunaan sesuatu perkataan dalam konteks tertentu, oleh itu, pemilihan perkataan atau istilah yang tidak tepat dalam pembinaan ayat, akan menjejaskan makna atau mesej yang ingin disampaikan. Jadi, satu laman web kamus yang dapat memaparkan medan tentang sesuatu perkataan seperti Entri, Definisi Entri, Peribahasa Entri, Nama Saintifik Entri dan sebagainya perlu dibangunkan. Informasi yang lengkap amat diperlukan kerana kepelbagaian makna dan sifat yang sering mengelirukan.

2. PENYATAAN MASALAH

Twitter adalah media sosial yang mempunyai pengguna melebihi 125 juta dan Twitter adalah salah satu platform sosial media yang digunakan di Malaysia. Twitter digunakan untuk memastikan followers mengetahui maklumat terkini sendiri ataupun bisnes.

Twitter dapat digunakan untuk komunikasi dengan followers nya (Collins, 2013). Semua pengguna Twitter dapat berkomunikasi dengan semua pengguna lain di mana-mana di dalam dunia ini, termasuk ahli keluarga dan kawan dapat berkomunikasi dengan cara yang senang dan dapat menghantar mesej dengan cepat dan lebih berkesan. Walau bagaimanapun, kebanyakan pengguna Twitter di Malaysia sering menggunakan pelbagai jenis singkatan dan bahasa perkhidmatan mesej ringkas (atau sms) dalam penghantaran mesej harian. Sebagai contoh, antara perkatan yang sering digunakan ialah seperti 4ever (selama-lamanya) atau adek (adik).

Mesej Twitter berbahasa Melayu mempunyai masalah yang sama seperti Bahasa Inggeris (Clerk & Araki, 2011) iaitu ia tidak mengikut peraturan ejaan, tatabahasa, tanda baca, salah ejaan, singkatan perkataan dan sebagainya (Goodman, 2001). Selain itu, Samsudin et al. (2012) telah mengenalpastikan Bahasa Melayu mempunyai 13 corak istilah hingar yang lazim. Bahasa Melayu juga bahasa yang ke-empat yang banyak digunapakai dalam aplikasi Twitter di seruluh dunia (Watanabe e tal., 2011).

3. OBJEKTIF KAJIAN

Matlamat utama adalah untuk pernormalan teks media sosial menggunnakan perdekatan kamus hybrid. Objektif projek ini adalah seperti berikut:

a) Mencadangkan kaedah menggunakan pendekatan model kamus untuk pernormalan teks ciap dalam bahasa Melayu.
b) Untuk membangunkan prototaip model kamus yang mengandungi singkatan perkataan yang mencukupi untuk menormalkan ciap Bahasa Melayu.
c) Membina set data ujian yang mengandungi ciap Bahasa Melayu untuk menguji keberkesanan kaedah yang dicadangkan.

4. METOD KAJIAN

Metodologi yang digunakan adalah bersifat eksperimen iaitu melibatkan pembangunan kaedah dan algoritma untuk pernormalan ciap Bahasa Melayu. Metodologi yang digunakan melibatkan beberapa fasa:

4.1 Fasa Keperluan (*Requirement)* - Fasa menentukan permasalahan dan objektif kajian. Fasa ini adalah untuk mengumpul keperluan kajian dan data. Fasa ini juga mengutamakan pengumpulan data dan menganalisis data. Data akan dikumpul dari Twitter dengan merekod setiap tweet secara manual. Kemudian, data akan dianalisis untuk mengenal pasti ciri-ciri dan objektif projek dan mengutamakan kerja-kerja pada ciri-ciri tersebut.

4.2 Fasa Rekabentuk (*Design*) - Fasa ini melibatkan menyediakan pelbagai rekabentuk sistem dan memilih reka bentuk sistem yang paling sesuai. Untuk mengenal pasti algoritma dan kaedah yang sesuai untuk proses pernormalan ciap Bahasa Melay. Mebangunkan set data untuk tujuan pengujian.

4.3 Fasa Pembangunan (*Development)* - Mebangunakan algoritma pernomalan ciap Bahasa Melayu. Fasa ini akan membincangkan aspek pembangunan dan implementasi model kamus yang dibina berdasarkan keperluan untuk semua ciap yang matlamat dinormalisasikan. Memastikan matlamat dan objektif pernormalan ciap Melayu

dicapai. Kod-kod yang dihasilkan akan disemak sama ada mencapai keperluan atau sebaliknya.

4.4 Fasa Pengujian (*Testing*) – Fasa ini akan menguji ketepatan proses penormalan yang dihasilkan. Kod-kod akan dikumpul semula bagi tujuan menguji. Kaedah penilaian pretasi algorithma pernormalan ciap akan diambil. Jika sebahagian kod mempunyai masalah, penambahbaikan semula perlu dilaksanakan.

5. HASIL KAJIAN

Bahagian ini menerangkan hasil dapatan selepas pernormalan teks media sosial yang telah dinormalkan dengan cara perdekatan kamus hybrid.

```
[ ]  import nltk
     import re
     nltk.download('punkt')


     my_dic ={'1': 'satu','3':'Tiga','4': 'empat','5': 'lima','6':'enam', '7':'tujuh','8':'Lapan', '9':'Sembilan', '10':'sepu
             'sbr': 'sabar', 'scr': 'secara', 'sdar': 'sedar', 'sdp': 'sedap','tk':'tak' ,'sdr': 'sedara', 'sg': 'sungai', '

     punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
     input_str="tunggu who ni mmg lambat nk umum pandemic pun lambat"                    // input string
     input_str =input_str.lower()                                                        // lower case
     result=re.sub("\d","",input_str)                                                    //remove punctuation
     no_punct = ""
     for char in input_str:
        if char not in punctuations:
            no_punct = no_punct + char
     print(no_punct)

     from nltk.tokenize import word_tokenize                                              // tokenize
     tokens= word_tokenize(no_punct)

     numoftokens = print(len(nltk.word_tokenize(input_str)))                              // total number of token
     print(tokens)
     output = " "
     for t in tokens:

       if t in my_dic:

         data = my_dic.get(t, "");

         output = output + data + " ";

       else :

         output = output + t + " ";

     print(output);
```

Rajah 1 : Rajah menunjukkan kod-kod implikasi dalam pernormalan teks media sosial Bahasa Melayu.

Bahagian ini akan memproses peratusan ketepatan untuk 200 ayat yang telah diproses dan dibahagi ke dalam 4 julat, iaitu (Bilangan ayat yang dapat 0 hingga 25 peratus ,

bilangan ayat yang dapat 26 hingga 50 peratus , bilangan ayat yang dapat 51 hingga 75 peratus , bilangan ayat yang dapat 76 hingga 100 peratus). Peratusan Ketepatan akan direkod dalam bentuk seperti berikut:

Peratusan Ketepatan(%)	Bilangan ayat	Peratusan ayat yang betul diproses dalam kategori (%)
0-25	10	5
26-50	27	13.5
51-75	64	32
76-100	99	49.5

Rajah 6.2: Rajah ini menunjukan bilangan ayat yang mendapat peratusan ketepatan masing-masing.

Peratusan Ketepatan adalah percentage yang didapat oleh setiap ayat dan dikategori kepada empat (0-25%, 26-50%, 51-75%, 76-100%).

Bilangan ayat menunjukkan berapa ayat yang mendapat peratusan ketepatan yang tertentu.

Peratus ayat yang betul diproses dalam kategori adalah peratus/ percentage ayat yang dalam kategori (peratusan ketepatan) tersebut.

Sebanyak 49.5 peratus ayat (99 ayat) daripada 200 ayat mendapat 76 hingga 100 peratus ketepatan dalam proses pernormalan yang menggunakan kamus hybrid yang dibina dalam proses pembangunan. Selain itu, 32 peratus (64 ayat) daripada 200 ayat mendapat peratus ketepatan dalam 51 hingga 75 peratus disebabkan perkataan yang diguna dalam singkatan perkataan yang tidak sering digunakan dalam sosial media seperti kja(kerja), sorg(seorang) dan truskn(teruskan). Manakala, 13.5 peratus (27 ayat) daripada 200 ayat mendapat peratusan ketepatan sebanyak 26 hingga 50 peratus, dan ayat yang mendapat keputusan ini disebabkan dialek yang digunakan seperti pon(pun), diorang (dia orang) dan apo(apa). Cuma 5 peratus (10) ayat yang mendapat 0 hingga 25 peratus ketepatan disebabkan tatabahasa seperti xpelik (tidak pelik), xyah (tak perlu), dan mnjaga(menjaga).

6. KESIMPULAN

Kajian ini mempunyai tiga objektif utama: - (1) Untuk mencadangkan senibina arkiteture baru yang menggunakan perdekatan model kamus untuk penormalan teks ciap dalam bahasa Melayu.(2) Untuk mereka prototaip model kamus yang mengandungi singkatan perkataan yang mengcukupi untuk menormalkan ciap Melayu. (3) Membina set data ujian yang mengadungi ciap Melayu. Dan kekangan yang masih perlu diselesaikan seperti fungsi predikasi perkataan diikuti dengan setiap perkataan dan fungsi untuk membetulkan tatabahasa diperlukan supaya perkataan seperti (sediakn) dan (mnbantu) dapat dinormalkan dengan betul. yang dinormalkan ke bahasa Melayu yang piawai. Seni bina penyelidikan ini akan disumbangkan oleh penyelidikan saya mengenai Pernormalan Teks Media Sosial Bahasa Melayu.

7. RUJUKAN

Nor Azlizawati Muhamad, Norisma Idris & Mohammad Arshi Saloot. 2017. Proposal: A Hybrid Dictionary Modelling Approach for Malay Tweet Normalization. Journal of Physics: Conference Series, IOP Conf. Series: Journal of Physics: Conf. Series 806 (2017) 012008

Clark, E. & Araki, K., 2011. Text normalization in social media : progress , problems and applications for a pre-processing system of casual English. , 27(Pacling), pp.2–11.

Basri, S.B., Alfred, R. & On, C.K., 2012. Automatic spell checker for Malay blog. In Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on. pp. 506–510.