Identifikasi Bahasa Daerah di Indonesia Dengan Multinomial Naïve Bayes

Maulana, Muhammad Ardhianda[1], Suryani, Arie Ardiyanti[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]ardhiandam@student.telkomuniversity.ac.id, [2]ardiyanti@telkomuniversity.ac.id

Abstrak

Saat ini telah banyak penelitian yang melakukan identifikasi bahasa, namun untuk identifikasi bahasa daerah di Indonesia belum banyak hasil yang diberikan. Untuk itu penelitian ini akan membahas tentang identifikasi bahasa daerah di Indonesia dengan menggunakan tujuh bahasa yaitu, Bahasa Indonesia, Jawa, Sunda, Minang, Muna, Bugis, dan Madura. Pendekatan yang digunakan untuk mengidentifikasi bahasa di penelitian ini menggunakan metode Multinomial Naïve Bayes. Pendekatan ini dilakukan menghitung probabilitas setiap pola kata maupun deretan kata yang muncul pada kalimat berlabel. Model probabilitas yang dihasilkan kemudian digunakan untuk menentukan kelas kalimat baru yang akan ditentukan bahasanya. Performansi metode identifikasi bahasa ini diukur dengan melakukan dua skenario pengujian. Pengujian pertama dilakukan untuk melihat pengaruh pola n-gram terhadap *F-measure*, sedangkan pengujian kedua dilakukan untuk mengobservasi pengaruh jumlah data latih terhadap *F-measure*. Hasil pengujian menunjukkan bahwa pola unigram dan bigram memberikan hasil akurasi tertinggi sebesar 98,86%. Adapun untuk jumlah data latih sebesar 1500 kalimat pada setiap bahasa menunjukkan akurasi sebesar 98%.

Kata kunci : identifikasi bahasa, bahasa daerah, multinomial naïve bayes

Abstract

Currently, there has been a lot of research that has carried out language identification, but not many results have been provided for identifying regional languages in Indonesia. For this reason, this research will discuss the identification of local languages in Indonesia using seven languages, namely, Indonesian, Javanese, Sundanese, Minang, Muna, Bugis and Madurese. The approach used to identify languages in this research uses the Multinomial Naïve Bayes method. This approach is used to calculate the probability of each word pattern or row of words appearing in a labeled sentence. The resulting probability model is then used to determine the class of new sentences for which the language will be determined. The performance of this language identification method is measured by conducting two test scenarios. The first test was to find out the effect of n-gram pattern on the F-measure, while the second test was to observe the effect of the amount of training data on the F-measure. The test results show that the unigram and bigram patterns provide the highest accuracy results of 98.86%. As for the amount of training data of 1500 sentences for each language, it shows an accuracy of 98%.

Keywords: language identification, local languages, multinomial naïve bayes

1. Pendahuluan

Latar Belakang

Identifikasi Bahasa (*Language Identification*) adalah salah satu *task* dalam Pemrosesan Bahasa Alami (*Natural Language Processing*) yang bertujuan mengenali sebuah bahasa yang ada di dalam sebuah teks atau dokumen.[1] Identifikasi Bahasa dapat diaplikasikan di berbagai area, proses, atau *task* seperti *document resume, text classification, machine translation,* dan lainnya. Salah satu contoh penggunaan dari sistem identifikasi bahasa adalah pengembangan lebih lanjut dari *chatbot* dimana sistem secara otomatis mengidentifikasi bahasa dari lawan bicara sebelum memberi respon, analisis sentimen pada kalimat.

Saat ini penelitian terhadap *Language Identification* mengalami perkembangan yang pesat. Terdapat beberapa hasil penelitian yang membahas identifikasi Bahasa diantaranya *labeling hate speech*,[2] dan mengidentifikasi bahasa isyarat.[3] Namun untuk bahasa dengan korpus yang terbatas atau bahasa yang minor digunakan dalam penelitian, relatif sulit untuk menemukan hasil penelitian yang membahas tentang identifikasi bahasa tersebut, salah satu bahasa yang dimaksud adalah bahasa daerah yang terdapat pada di Indonesia, adapun saat ini penelitian identifikasi bahasa terhadap bahasa daerah yang dilakukan adalah *spoken language identification* dan identifikasi bahasa daerah di Indonesia dengan berbagai metode *machine learning*.

Topik dan Batasannya