INTISARI

Analisis Klasifikasi Topik Menggunakan Metode *Naïve Bayes Classifier*, *Naïve Bayes Multinomial Classifier*, dan *Maximum Entropy* pada Artikel Berita

Oleh

Nurul Masithoh
12/334680/PA/14913

Pada era pesatnya informasi yang berbasis *website* memberikan kemudahan dalam memperoleh informasi. Kebutuhan informasi yang cepat, akurat, dan berasal dari berbagai topik isu kehidupan telah menjadi gaya hidup. Informasi sebagian besar tersedia dalam bentuk data tekstual yang tidak memiliki pola terstruktur. Ilmu statistika semakin diaplikasikan pada berbagai bentuk data, salah satunya data tekstual. *Text mining* adalah analisis yang berperan dalam pengolahan data tekstual. Analisis *text mining* yang sering digunakan adalah klasifikasi teks atau klasifikasi topik. Dari hasil klasifikasi topik dapat disimpulkan jenis topik dari suatu dokumen yang tidak terstruktur dengan melihat karateristik dokumen. Metode yang akan dibahas kali ini adalah metode *Naïve Bayes Classifier*, *Naïve Bayes Multinomial Classifier*, dan *Maximum Entropy* untuk menentukan kelas topik suatu dokumen secara multikelas yaitu kelas topik bola, ekonomi, kesehatan, dan travel.

Naïve Bayes Classifier dan *Naïve Bayes Multinomial Classifier* merupakan metode klasifikasi menggunakan aturan Bayesian memanfaatkan probabilitas prior dan probabilitas bersyarat dari kemunculan kata, untuk *Naïve Bayes Classifier* dan frekuensi kemunculan kata untuk *Naïve Bayes Multinomial Classifier* pada masing-masing kelas dokumen *training*. Sedangkan metode *Maximum Entropy* menggunakan rata-rata informasi yang terkandung dalam dokumen yang dimaksimalkan sehingga semakin memberikan hasil klasifikasi yang akurat. Nilai tersebut akan digunakan untuk menentukan kelas topik suatu dokumen *testing* dengan melihat nilai *maximum a posterior* masing-masing kelas. Data yang telah diklasifikasikan kemudian dihitung tingkat akurasi kebenarannya. Dari perbandingan rata-rata nilai akurasi klasifikasi dengan ketiga metode di atas, diperoleh urutan yang paling tinggi menggunakan metode *Maximum Entropy* sebesar 99,31%, *Naïve Bayes Classifier* sebesar 98,82%, dan yang paling rendah yaitu dengan metode *Naïve Bayes Multinomial Classifier* sebesar 97,39%. Percobaan ini dilakukan pada 1440 artikel berita dimana setiap metodenya dilakukan 87 percobaan. Selain membandingkan ketiga metode, dilakukan analisis tentang aspek pengaruh penggunaan jumlah token, jumlah data, dan keseragaman jumlah data pada nilai akurasi. Diperoleh kesimpulan jika ketiga aspek kurang mempengaruhi nilai akurasi di setiap metode.

Kata kunci : Klasifikasi topik, *text mining*, aturan Bayesian, *Naïve Bayes Classifier, Naïve Bayes Multinomial Classifier, Maximum Entropy*.

ABSTRACT

Topics Classification Analysis Using *Naïve Bayes Classifier, Naïve Bayes Multinomial Classifier,* and *Maximum Entropy* for News Articles

by

Nurul Masithoh

12/334680/PA/14913

The rapid growth of website based communication nowdays gives easier access to obtain more information. The demand for better accuary and faster information is increasing now due to the modern lifestyle. Most of the data are provided in textual verse aren't presented in any structured patterns. Statistics is playing a prominent role in this part, especially in completing textual data. *Text mining* is an analitical system that used in textual data. This method which are often applied in analytical process are text classification and topic classification. Based on result of topic classification, analyst can decide the topic type. The topic itself is concluded based on the document characteristic. This paper will discuss about *Naïve Bayes Classifier*, *Naïve Bayes Multinomial Classifier*, dan *Maximum Entropy*. Those methode will be applied to define the topic class based on multiclasses: ball topic class, health, economy, and tavel.

Naïve Bayes Classifier and *Naïve Byaes Multinomial Classifier* are classifying methods that use Bayesian rules. The rules are including prior probability and conditional probability based on the appearance of words, for *Naïve Bayes Classifier* and the frequency of words appearance for *Naïve Bayes Multinomial Classifier* in each document *training*. While *Maximum Entropy* uses approximate information which is included in the document. This information is optimized so that the classification becomes more accurate. That value will be applied to decide the topic class, by looking at the maximum a posterior in each class. Classified data would be counted in order to find the accuracy level. Based on the comparision of approximate values from those methods, the leveling of accuracy. The highest level is *Maximum Entropy* (99,31%), then *Naïve Bayes Classifier* (98,82%), and the lowest level is *Naïve Bayes Multinomial Classifier* (97,39%). This research was applied on 1440 news articles. Each articles underwent 87 trials. Not only comparing those three methods, but also analyzing the spect of token data usage, number of data usage, and the similary number of data on value accuary. The conclusion is those three aspects don't have influence the data accuary in each method.

Keywords : Topic classification, *text mining*, Bayesian rules, *Naïve Bayes Classifier*, *Naïve Bayes Multinomial Classifier*, *Maximum Entropy*.