**BAB III**

**PEMBAHASAN**

**3.1 Pengertian dan Tujuan Meta ToolKit**

Meta adalah ToolKit atau perangkat lunak gratis dan open source (sumber terbuka) yang dapat digunakan untuk menganalisis data teks.

Tujuan META adalah untuk memperbaiki dan melengkapi perangkat lunak machine learning dan pengambilan informasi open source (sumber terbuka) saat ini. Lingkungan yang ada dari perangkat lunak open source (sumber terbuka ) ini cenderung cukup terfragmentasi. META Secara khusus menyediakan kerangka kerja pemersatu untuk pengindeksan teks dan metode analisis, yang memungkinkan pengguna untuk menjalankan eksperimen terkontrol dengan cepat. Ini memodularisasi pembuatan fitur, representasi instance, format penyimpanan data, dan implementasi algoritma; hal ini memungkinkan peneliti dan siswa untuk melakukan transisi tanpa batas selama dimensi ini dengan sedikit usaha. Modularitas META mendukung eksplorasi, mendorong kontribusi, dan meningkatkan visibilitas ke bagian dalamnya. Fakta-fakta ini menjadikannya sebagai perangkat pendamping yang sempurna untuk analisis data teks (Zhai & Massung, 2016).

**3.2 Arsitektur (Architecture)**

Semua data yang diproses di META disimpan dalam indeks. Ada dua jenis indeks: forward\_index dan inverted\_index. Yang pertama dikunci oleh ID dokumen, dan yang kedua dikunci oleh ID istilah

* forward\_index digunakan untuk aplikasi seperti pemodelan topik dan sebagian besar tugas klasifikasi.
* inverted\_index digunakan untuk membuat mesin pencari, atau melakukan klasifikasi dengan knearest-neighbour atau algoritma serupa.

Karena setiap aplikasi META mengambil indeks sebagai input, semua data yang diproses dapat dipertukarkan di antara semua komponen. Ini juga memberikan keuntungan besar untuk klasifikasi. META mendukung klasifikasi out-of-core secara default! Jika set data cukup kecil (seperti asumsi kebanyakan toolkit lain), cache dapat digunakan seperti no\_ evict\_cache untuk menyimpan semuanya dalam memori tanpa mengorbankan kecepatan apa pun.

Ada empat format masukan korpus, yaitu:

line\_corpus. setiap dataset terdiri dari satu hingga tiga file: corpusname.dat. setiap dokumen muncul dalam satu baris corpusname.dat.labels. file opsional yang