BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Jurnal adalah sebauh hasil laporan peneliti yang telah dilakukan secara ilmiah, sebagian besar jurnal penelitian dapat dipertanggungjawabkan keilmiahannya tergantung dari metode yang digunakan dalam membuat dan pnyusunan laporan jurnal penelitian tersebut. Karena beragam jurnal penelitian, jurnal tersebut dikelompokkan berdasarkan subjeknya masing-masing diantaranya adalah agama, social, insudtri, teknologi, sains, dll. Pada setiap jurnal, diketahui bahwa penyusunan jurnal tersebut akan dikategorikan berdasarkan setiap subjeknya untuk mempermudah pembaca dalam memilih subjek yang diinginkan. Dari permasalahan ini penulis melakukan penelitian tentang sebuah system yang dapat melakukan klasifikasi secara otomatis dengan menggunakan abstrak pada jurnal dengan metode *Term Frequency Inverse Document Frequency* dan *Multinomial Naïve Bayes*.

Naïve Bayes Classifier atau disingkat NBC adalah salah satu metode klasifikasi yang berakar unutk melakukan pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu class. Bayesian classification didasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengan decesion tree dan neural network. Bayesian classification terbukti memiliki akurasai dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar [1]. Metode tersebut merupakan pendekatan statistic untuk melakukan persoalan klasifikasi, kemudian menggunakan *Naïve Bayes* ini untuk melakukan klasifikasi dalam *text mining*. Pada kasus probabilitas dari suatu data [3]. Pada kasus [2] metode NBC digunakan mengklasifikasikan artikel berita, data yang diteliti adalah data artikel berita berbahasa Indonesia yang diambil dari web portal. Pada kasus [3], NBC digunakan untuk mengklasifikasikan dokumen dengan konten E-government dengan data berupa dokumen dengan format HTML yang diubah menjadi TXT.

Term Frequency Inverse Document Frequency atau TF-IDF adalah metode pembobotan kata yang bertujuan untuk memberikan bobot nilai pada setiap kata. Pada kasus [8], klasifikasi menggunakan metode *Naïve Bayes* menggunakan hasil dari TF-IDF memberikan hasil rata-rata akurasi yang lebih baik dibandingkan tanpa menggunkan TF-IDF[4].

Penelitian menggunakan metode NBC banyak ditemui pada klasidikasi dokumen berita[2], peneliti akan melakukan klasifikasi abstrak pada jurnal menggunakan metode Naïve Bayes Classifier dan Term Frequency Inverse Document Frequency dengan harapan yang akan didapatkan dapat mempermudah pembaca dalam menentukan subjek jurnal yang akna dibaca. Abstrak jurnal pada penelitian ini menggunakan 8 subjek yaitu Agama, Agrikultur, Ekonomi, Industri, Kesehatan, Pendidikan, Sains, dan Sosial. Dengan masing-masing data tiap subjek adalah 100 abstrak jadi total keseluruhan adalah 800 abstrak yang akan dilakukan klasifikasi. Peneliti memilih 8 subjek tersebut karena abstrak tersebut memiliki ketertarikan satu sama lain yang mana tiap kelas/subjek memiliki kata yang sama tetapi subjeknya berbeda.

Komputer tidak dibekali kemampuan untuk memahami bahasa manusia secara langsung, sehingga komputer tidak bisa membedakan antara gabungan kata yang membentuk frasa dan yang bukan. Salah satu cara untuk mengetahui perbedaan frasa dan bukan frasa, sebelumnya komputer harus mengetahui ciri-ciri frasa. Supaya komputer mengetahui ciri- ciri frasa, komputer harus mampu memberi label tiap kata sesuai kelas katanya. Proses pemberian kelas kata ini dikenal dengan nama POS-Tagger. Hal ini, karena proses POS-Tagger bisa dipandang sebagai proses klasifikasi suatu rangkaian atau urutan tag untuk tiap kata dalam suatu kalimat. Kelebihan dari proses POS-Tagger menggunakan pendekatan probabilistik adalah adanya proses training dalam pemberian kelas kata, sehingga tidak tergantung pada aturan kelas kata [5].

Pada penelitian tugas akhir ini membahas tentang klasifikasi text. Data yang digunakan menggunakan data abstrak pada jurnal yang di ambil dari salah satu *website* http://sinta.ristekbrin.go.id/ , kemudia data tersebut akan diolah melalui proses *preprocessing,* ekstraksi fitur, dan klasifikasi. *Output* pada penelitian tugas akhir ini berupa akurasi dan *F1-Score* untuk mengukur performa klasifikasi. Penelitian tugas akhir ini memiliki dua rumusan masalah yaitu, mengetahui kinerja Algoritma Multinomial Naïve Bayes dalam melakukan klasifikasi dokumen pada abstrak dan bagaimana hasil penerapan Algoritma Multinomial Naïve Bayes dengan menggunakan *Part of Speech* (POS) *Tagger* atau tanpa menggunakan POS Tagger.

Penelitian ini menggunakan ekstraksi fitur POS Tagger untuk menentukan POS tiap kata dalam dokumen dan fitur *Term Frequency Inverse Document Frequency* atau TF-IDF dengan metode

Multinomial Naïve Bayes. Fitur yang digunakan adalah fitur uni-gram yang mana menggunakan satu suku kata yang akan dilakukan pembobotan fitur sebelum dilakukannya klasifikasi.

Untuk mengetahui pengaruh *preprocessing* tanpa *stemming*, *stopword*, dan dengan menggunakan *stemming* dan *stopwords.*