

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Informasi didefinisikan sebagai data yang telah melalui proses sedemikian rupa sehingga dapat meningkatkan pengetahuan orang yang menggunakan data tersebut (Alshikhi and Abdullah, 2018). Informasi bisa didapatkan dari berbagai bidang keilmuan, salah satunya dalam bidang kesehatan. Dalam bidang kesehatan, artikel ilmiah kesehatan adalah salah satu sumber yang dapat memberikan sebuah informasi tentang penelitian baru, yang didukung oleh data yang relevan serta kuat dan dapat dibuktikan keandalannya (Marusic and Marusic, 2009). Artikel ilmiah kesehatan salah satu sumber informasi untuk mendapatkan pengetahuan klinis paling baru yang membahas tentang tanda dan gejala pasien, diagnosis, dan prosedur medis. Untuk membantu pembaca mendapatkan informasi terbaru secara cepat, perlu dilakukan ekstraksi entitas dalam sebuah artikel ilmiah kesehatan. Maka dari itu diperlukannya sebuah sistem yang mampu mengidentifikasi dan mengklasifikasikan entitas dari sebuah artikel ilmiah kesehatan secara otomatis untuk memudahkan seorang individu untuk menemukan informasi tertentu.

Ekstraksi informasi merupakan sebuah metode untuk menghasilkan informasi secara spesifik dari sebuah teks atau dokumen (Arifandi, 2017). Ekstraksi informasi memiliki struktur bahasa dan entitas kata, untuk mengidentifikasi entitas suatu kata pada kalimat diperlukan proses *Named Entity Recognition* (NER). Pengidentifikasian entitas suatu kata secara umum merujuk pada orang, tempat, organisasi, dan perusahaan (Mansouri and Affendey, 2008). Secara khusus NER dapat diimplementasikan ke dalam bidang kesehatan untuk mengidentifikasi entitas berupa nama penyakit, gejala, nama obat, dan lainnya yang berhubungan dengan kesehatan. Namun penerapan sistem NER untuk kesehatan lebih sulit dibandingkan dengan penerapan pada NER umum, dikarenakan lebih kompleks (Yao *et al.*, 2015). Penerapan sistem NER untuk kesehatan yang cukup kompleks ini mendorong para peneliti untuk melakukan penelitian terkait dengan sistem NER untuk kesehatan. Beberapa penelitian mengenai sistem NER untuk kesehatan yang sudah pernah dilakukan.

Disease Named Entity Recognition Using Conditional Random Fields yang diteliti oleh (Ur Rahman, Hahn and Segall, 2016), mereka melakukan penelitian menggunakan metode CRF untuk klasifikasi nama penyakit. Penelitian mereka

menggunakan *corpus* NCBI yang terdiri dari 793 abstrak dengan total 2783 kalimat dan 6900 nama penyakit. Penelitian mereka mendapatkan hasil tertinggi dengan *precision* sebesar 94%, *recall* sebesar 94%, dan *f-measure* sebesar 94% penggunaan fitur *POS-Tagging* meningkatkan nilai *precision* dari 77% menjadi 87%, *recall* dari 76% menjadi 74%, dan *f-measure* dari 74% menjadi 86.

Medical Entity Recognition using Conditional Random Field (CRF) yang diteliti oleh (Herwando, Jiwanggi and Adriani, 2017) mengklasifikasikan 4 entitas di dalam bidang medis yaitu penyakit, gejala, terapi, dan obat. Mereka menggunakan *corpus* yang didapatkan dari postingan forum *website* kesehatan yaitu TanyaDok, DokterSehat, KlikDokter, DetikHealth, dan AloDokter. Mereka melakukan eksperimen dengan menggabungkan dari 15 fitur untuk dapat mencari hasil yang terbaik dan didapatkan hasil akurasi tertinggi dengan nilai *precision* sebesar 70,97%, *recall* sebesar 57,83% dan *f-measures* sebesar 63,69%. Entitas dengan hasil akurasi tertinggi adalah entitas penyakit dengan nilai *precision* sebesar 73,30%, *recall* sebesar 63,34% dan *f-measures* sebesar 67,84%. Sedangkan untuk entitas dengan hasil akurasi terendah adalah entitas obat dengan nilai nilai *precision* sebesar 68,97%, *recall* sebesar 45,37% dan *f-measures* sebesar 54,05%. Hal ini disebabkan entitas obat memiliki jumlah data yang lebih sedikit dibandingkan dengan entitas lain.

Berdasarkan latar belakang tersebut, penelitian ini mengusulkan NER untuk mengidentifikasi dan mengklasifikasi entitas dari artikel ilmiah kesehatan yang didapatkan dari Jurnal Penyakit Dalam Indonesia yang diterbitkan oleh Departemen Ilmu Penyakit Dalam Fakultas Kedokteran Universitas Indonesia. Dikarenakan belum adanya penelitian yang mengidentifikasi entitas dari artikel ilmiah kesehatan tersebut, dan dari analisis yang didapat berdasarkan beberapa penelitian di atas, metode yang akan digunakan adalah CRF. Metode CRF memiliki banyak kelebihan dibandingkan dengan model probabilitas yang lain seperti *Maximum Entropy Markov Model* (MEMM) dan *Hidden Markov Model* (HMM) dan CRF mempunyai tingkat akurasi yang paling baik diantara semua metode yang digunakan, dan diharapkan dalam penelitian ini mendapatkan hasil yang lebih akurat.

Selain mengidentifikasi dan mengklasifikasi entitas dari artikel ilmiah kesehatan menggunakan metode CRF, penelitian ini juga akan meninjau dari perspektif agama Islam mengenai penerapan NER pada artikel ilmiah kesehatan

menggunakan metode CRF, dilihat dari aspek manfaat dan boleh atau tidaknya dalam agama Islam.

Salah satu manfaat penerapan NER ini adalah dapat memudahkan pencarian entitas tertentu dalam artikel ilmiah kesehatan. Islam merupakan agama yang memberikan kemudahan dalam hidup, kemudahan merupakan salah satu prinsip penting dalam Islam. Allah SWT menghendaki manusia untuk memperoleh kemudahan, hal ini sebagaimana dinyatakan dalam firman-Nya:

... يُرِيدُ اللَّهُ بِكُمُ الْيُسْرَ وَلَا يُرِيدُ بِكُمُ الْعُسْرَ ... ﴿١٨٥﴾

Artinya: ... Allah menghendaki kemudahan bagimu, dan tidak menghendaki kesukaran bagimu ... (Q.S Al-Baqarah (2): 185).

Melihat adanya manfaat yang bisa diambil dari penelitian ini dan diperkuat dengan penjelasan dari ayat di atas, maka ayat tersebut dapat dijadikan sebagai pegangan untuk menyimpulkan apakah penerapan NER pada artikel ilmiah kesehatan menggunakan metode CRF diperbolehkan dalam agama Islam.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang di atas, maka diperoleh rumusan masalah sebagai berikut:

- a) Pencarian informasi tertentu pada artikel ilmiah kesehatan dapat memakan cukup waktu dikarenakan harus membaca setidaknya beberapa paragraf.
- b) Masih sedikit implementasi sistem NER untuk mengklasifikasikan dan mengidentifikasi artikel ilmiah kesehatan khususnya berbahasa Indonesia.

1.3 Rumusan Masalah

Berdasarkan penjabaran latar belakang di atas, maka diperoleh rumusan masalah sebagai berikut:

- a) Bagaimana cara melakukan ekstraksi fitur pada artikel ilmiah kesehatan?
- b) Bagaimana mengidentifikasi dan mengklasifikasikan entitas setiap kata pada artikel ilmiah kesehatan?
- c) Bagaimana mengimplementasikan metode CRF untuk sistem NER pada artikel ilmiah kesehatan?
- d) Bagaimana mengukur kinerja dari sistem NER pada artikel ilmiah kesehatan?
- e) Bagaimana tinjauan agama Islam dalam penerapan NER pada artikel ilmiah kesehatan menggunakan metode CRF?

1.4 Batasan Masalah

Adapun batasan masalah yang terdapat pada penelitian ini adalah sebagai berikut:

- a) Data *training* dan *testing* yang digunakan dalam penelitian ini adalah *corpus* yang didapatkan dari 29 artikel ilmiah kesehatan pada Jurnal Penyakit Dalam Indonesia. Jurnal tersebut dapat diakses secara *online* pada situs <http://jurnalpenyakitdalam.ui.ac.id/>.
- b) Sistem NER yang akan dibangun berbasis *stand alone* atau *desktop*, dan menggunakan bahasa pemrograman *python*.
- c) Entitas yang akan diklasifikasikan dalam penelitian ini berjumlah 7, yaitu organ, *sign&symptom*, pemeriksaan fisik, pemeriksaan penunjang, penyakit, terapi farmakologi, dan terapi non farmakologi.
- d) Format *corpus* yang digunakan dalam penelitian ini adalah format IOB2 (*Inside, Outside, Begin*).
- e) Metode evaluasi untuk mencari nilai akurasi yang digunakan adalah *precision, recall, dan f-measure*.

1.5 Tujuan dan Manfaat Penelitian

Adapun tujuan dari perancangan sistem NER ini adalah:

- a) Menghasilkan *corpus* yang didapat dari gabungan beberapa artikel ilmiah kesehatan.
- b) Melakukan ekstraksi fitur pada artikel ilmiah kesehatan.
- c) Melakukan identifikasi dan klasifikasi entitas setiap kata pada artikel ilmiah kesehatan.
- d) Mengevaluasi penggunaan algoritma CRF dalam pengklasifikasian entitas setiap kata pada artikel ilmiah kesehatan.
- e) Menghasilkan sebuah *prototype* yang berbentuk GUI sederhana.
- f) Meninjau implementasi klasifikasi entitas setiap kata pada artikel ilmiah kesehatan menggunakan algoritma CRF menurut agama Islam

Adapun manfaat dari perancangan sistem NER ini adalah:

- a) Membantu dunia kesehatan dalam pengklasifikasian entitas kata pada artikel ilmiah kesehatan.
- b) Penelitian ini dapat digunakan sebagai evaluasi untuk penelitian sejenis.
- c) Penelitian ini dapat digunakan sebagai *pre-processing* untuk penelitian NLP lainnya.