BAB 2 LANDASAN KEPUSTAKAAN

2.1 Kajian Pustaka

Kajian pustaka yang digunakan pada penelitian ini diantaranya penelitian yang telah dilakukan oleh Zulen dan Purwarianti pada tahun 2011, pada penelitian tersebut digunakan metode *EAT* untuk mengelompokkan jenis pertanyaan sehingga didapat *expected answer.* Kemudian jawaban akan dikelompokkan berdasar jenisnya apakah termasuk *Factoid atau Non-Factoid.* Langkah selanjutnya dalam menemukan jawaban apabila termasuk *Factoid* maka digunakan metode NER, yaitu menemukan jawaban dengan mencocokkan jenis entitas yang merupakan kandidat jawaban dengan mempertimbangkan jarak antara *clue word* dengan kandidat jawaban sedangkan jika *Non-Factoid* maka digunakan metode *Pattern Matching*, yaitu dengan cara mencocokkan *pattern* tiap kalimat dengan memperhitungkan jarak dan *pattern* dengan prioritas tertinggi (Zulen & Purwarianti, 2011). Pada penelitian tersebut didapatkan MRR sebesar 0,6191 untuk pertanyaan bertipe *factoid.*

Sedangkan, pada penelitian yang dilakukan Wang pada tahun 2016. Dalam menentukan jawaban dari pertanyaan yang diberikan pengguna adalah dengan menghitung jarak antar kata pada pertanyaan dengan kalimat yang ada dalam dokumen dengan mengadopsi algoritme *Djikstra* memetakan kata tersebut dalam *categori space* dan kemudian memilih top N kalimat yang terdekat dengan pertanyaan (Wang, et al., 2016). Pada penelitian tersebut didapatkan nilai *precision* antara 40% sampai dengan 60%.

Dalam pencarian *named entity* pada penelitian ini digunakan metode *Naive Bayes* dikarenakan pada penelitian sebelumnya yang dilakukan oleh Mahalaksi G.S. dalam mengklasifikasikan *named entity* berbasis dokumen dengan *Naive Bayes* diperoleh akurasi sebesar 79% (G.S., et al., 2016), sehingga dapat dikatakan bahwa metode *Naive Bayes* cukup baik untuk digunakan sebagai metode klasifikasi *named entity.*

Pada penelitian yang dilakukan oleh Gautama dibandingkan perhitungan jarak antara metode *Manhattan* dengan *Euclidean*, hasilnya diperoleh kesimpulan bahwa perhitungan menggunakan metode *Manhattan* memiliki waktu eksekusi lebih singkat dibandingkan dengan *Euclidean* (Gautama, et al., 2015). Sehingga pada penelitian ini digunakan metode *Manhattan* sebagai metode perhitungan jarak antara kandidat jawaban dengan kata kunci pertanyaan yang ada.

Pada penelitian ini masukan yang diberikan oleh pengguna berupa pertanyaan mengenai cerita rakyat Indonesia. Kemudian pertanyaan tersebut akan melalui tahap *preprosessing* dan selanjutnya adalah penentuan EAT untuk menentukan jenis *tag* dari kandidat jawaban yang akan dipilih. Kemudian penentuan kandidat jawaban dilakukan melalui metode *Named Entity Recognition* (NER). Kemudian menghasilkan keluaran berupa jawaban dari pertanyaan yang diberikan oleh pengguna.

2.2 Pemrosesan Bahasa Alami (*Natural Language Processing)*

Bahasa alami adalah bahasa yang telah berevolusi secara alami dan digunakan oleh manusia untuk tujuan komunikasi, misalnya Hindi, Inggris, Prancis, Jerman merupakan contoh dari bahasa alami. Bahasa alami juga disebut sebagai *Computational Linguistic* merupakan studi ilmiah bahasa dari perspektif komputasi. *Natural Language Processing* adalah bidang komputer dan linguistik manusia yang berkaitan dengan interaksi antara komputer dan bahasa manusia (Kumar, 2010).

2.3 Temu Kembali Informasi (*Information Retrieval)*

Sistem temu kembali informasi merupakan teknologi untuk mencari informasi yang sesuai dengan apa yang diinginkan pengguna dari berbagai sumber dan menampilkan semua hasil yang sesuai dengan yang diinginkan oleh pengguna (Mansouri, et al., 2008).

Selain itu sistem temu kembali informasi melibatkan data yang tidak terstruktur dan terstruktur yang arti semantiknya tidak didefinisikan dan hasil *query* yang diberikan berupa hasil yang berperingkat.

2.4 Named Entity Recognition

Named Entity Recognition merupakan komponen dari *question answering system*. *Named Entity Recognition* merupakan komponen dari sistem ekstraksi informasi dan merupakan metode yang dapat digunakan dalam *question answering system*. *Named Entity Recognition* merupakan bagian dari ekstraksi informasi yang mencari dan mengklasifikasikan kata dalam teks yang berupa kata maupun frasa kedalam kategori yang ada seperti lokasi, organisasi, nama dan sebagainya. Di Indonesia penggunaan NER dilakukan dengan mengkombinasikan kontekstual, *morphological* dan *part of speech feature* kedalam *knowledge engineering* (Wongso, et al., 2016)

2.5 Question Answering System

Question Answering System mempunyai kaitan yang erat dengan ekstraksi informasi. Sebelum mendapat jawaban dari *query* tertentu, sistem perlu mengekstrak informasinya. Pertanyaan yang dimasukkan ke sistem dapat berupa pertanyaan yang panjang dan kompleks, atau mungkin dalam bentuk yang sederhana. Namun kita harus mengambil informasi dari pertanyaan tersebut, selanjutnya jawaban yang diberikan kepada pengguna harus sesingkat mungkin. Jawaban tersebut harus berisi jawaban yang tepat yang menggambarkan kebutuhan pengguna (Wongso, et al., 2016).

Dalam proses untuk menemukan jawaban *question answering system* memiliki tiga tahapan proses yang harus dilalui yaitu *question analyzer, passage retriever* dan *answer finder* untuk lebih jelasnya akan dijelaskan pada subbab selanjutnya.

2.5.1 *Question Annalyzer*

Question Analyzer merupakan tahapan dimana pertanyaan yang dimasukkan oleh pengguna di proses dengan menggunakan *Expected Answer Type* (EAT) dengan mempertimbangkan kata kunci pertanyaan dan kata kunci pendukung dari pertanyaan yang diberikan oleh pengguna (Zulen & Purwarianti, 2011).

1. Klasifikasi EAT pada pertanyaan yang diberikan oleh pengguna
   Klasifikasi EAT terhadap pertanyaan yang diberikan pengguna dilakukan dengan melihat kata kunci pertannyaan dan kata kunci pendukung. Adapun klasifikasi EAT dapat dilihat pada Tabel 2.1.

Tabel 2.1 Klasifikasi Expected Answer Type

EAT	Question Words	Clue Words
Factoid		
Person	Siapa, Siapakah	-
Location	Di mana , Di manakah	-
	Ke mana , Ke manakah	-
	Dari mana, Dari manakah	-
Date/Time	Kapan, Kapankah	-
	Berapa, Berapakah	Tanggal, bulan, tahun, abad, jam, menit detik
Organization	Apa, Apakah	Organisasi, perusahaan, badan, institusi, lembaga, partai, komisi, sekolah, komite, universitas
Quantity	Berapa, Berapakah	-
Non Factoid		
Definition	Apa, Apakah	Definisi, yang dimaksud, pengertian, arti
Reason	Mengapa, Kenapa	-
	Apa, Apakah	Penyebab, menyebabkan
Method	Bagaimana, Bagaimanakah	-

Sumber: (Zulen & Purwarianti, 2011)

2. *Keyword Extraction*
   Keyword extraction dilakukan dengan melihat POS tertentu sebagai *keywords* (Purwarianti *et al.*, 2007). Pada Tabel 2.2, *tagger* POS yang digunakan adalah *IPOSTagger*(Wicaksono dan Purwarianti, 2010). POS *Tag* yang akan diambil sebagai kata kunci ditunjukkan pada Tabel 2.2.

Tabel 2.2 Aturan *POSTagging*

POS	POS Name	POS	POS Name
NN	Common Noun	CDO	Ordinal Numerals
NNP	Proper Noun	CDC	Collective Numerals
NNG	Genitive Noun	CDP	Primary Numerals
VBI	Intransitive Verb	JJ	Adjective
VBT	Transitive Verb	FW	Foreign Words

Sumber: (Wicaksono dan Purwarianti, 2010)

2.5.2 *Passage Retriever*

1. Koleksi Pustaka (Korpus)

   Korpus digunakan sebagai sumber untuk mencari jawaban dari pertanyaan yang diberikan oleh pengguna. Korpus harus dalam bentuk dokumen dalam bahasa Indonesia dan harus termasuk dalam satu topik yang sama. Terdapat dua macam korpus berdasarkan cara mengaksesnya yaitu korpus *offline*, dapat berupa *e-book* maupun database *dumb* yang berasal dari artikel bahasa Indonesia dari Wikipedia dapat pula berupa korpus *online* yang berupa koleksi artikel dari sebuah *website*.

2. Teknik Pencarian

   Pencarian dokumen dilakukan melalui kata kunci yang dihasilkan dari *question analyzer*. Pencarian dilakukan dengan mencari dokumen dari korpus yang mengandung kata kunci dan kemudian mencari paragraf dari dokumen-dokumen yang mengandung kata kunci tersebut yang akan dijadikan masukan dalam *answer finder*.

2.5.3 *Answer Finder*

1. *Factoid Question*

   Factoid question sering dianggap setara dengan pencarian informasi. Dalam *factoid question answering* diberikan *knowledgebase* dan sebuah *query* yang digunakan dalam menemukan jawaban dari *factoid question* tersebut (Iyyer, et al., 2014).

   Metode yang dapat digunakan pada *Factoid Question* adalah *machine learning* dan *Named Entity Recognition* (NER) yang digunakan untuk menemukan kandidat jawaban dengan mengekstraksi *named entity* pada dokumen atau paragraf untuk mendapatkan kandidat jawaban yang memiliki Nes yang sesuai dengan *expected answer type* (EAT) dari kata kunci pertanyaan yang diberikan oleh pengguna. Dalam *Factoid Question* contoh EAT yang dapat digunakan adalah *Person, Organization, Location, Datetime,* dan *Quantity* yang dapat diekstrak menggunakan NE *tagger*.

2. *Non-Factoid Question*

   Metode yang dapat digunakan pada *Non-Factoid Question* adalah *Pattern Matching* dan *Semantic Analysis* (Zulen & Purwarianti, 2011). Dalam metode *Pattern Matching, answer finder* menjawab menggunakan aturan yang mempertimbangkan *surface expression* (pola kalimat) dan *linguistic clue (clue words)* dari kalimat. Dalam metode analisis semantik jawaban diperoleh dengan menyatukan representasi pertanyaan dengan fakta yang diketahui.

2.6 Naive Bayes

Naive Bayes merupakan pengklasifikasi sederhana berdasarkan penerapan Teorema Bayes dengan asumsi independen. Klasifikasi berdasarkan asumsi merupakan prinsip hipotesis posterior maksimal untuk mengidentifikasi benda

yang paling mungkin tergolong dalam kategori. Teorema Bayes menunjukkan hubungan antara satu kondisional probabilitas dan kebalikannya (An, et al., 2017).

Dalam pengelompokan *Naive Bayes*, digunakan persamaan untuk menghitung probabilitas masing-masing kelas A yang diberikan nilai B_i dari semua atribut untuk contoh yang harus diklasifikasikan. Persamaan yang dapat digunakan dalam klasifikasi *naive bayes* dapat dilihat pada Persamaan 2.1.

$$P(A|B_1 \dots B_n) = P(A) \prod_i \frac{P(A|B_i)}{P(A)} \quad (2.1)$$

2.7 Pengujian

2.7.1 *Confusion Matrix*

Confusion Matrix merupakan metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. *Confusion Matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Untuk menghitung *Confusion Matrix* dapat digunakan Tabel 2.3.

Tabel 2.3 *Confusion Matrix*

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	*True Positive*	*False Negative*
Negatif	*False Positive*	*True Negative*

True positives adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false* negatives adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai *negative*, kemudian masukkan data uji. Setelah data uji dimasukkan ke dalam *confusion matrix*, hitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *sensitivity(recall), specificity, precision* dan *accuracy* (Leidiyana, 2013).

2.7.2 *Precision*

Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Perhitungan tingkat ketepatan dokumen dengan menggunakan precision dapat dilakukan dengan menggunakan Persamaan 2.2.

$$P = \frac{|relevan \cap dokumen\ diperoleh|}{|dokumen\ diperoleh|} \quad (2.2)$$

2.7.3 Precision Pada Peringkat K

Precision pada peringkat ke *k* atau biasa disebut dengan *precision@k* merupakan perhitungan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem dengan menghitung persentase jawaban sejumlah *k* teratas.

Precision pada peringkat ke *k* didapatkan dengan menentukan *k* peringkat sebagai *threshold,* kemudian menghitung persentase jawaban relevan sebanyak

k peringkat dan mengabaikan jawaban dengan peringkat lebih dari *k* (Stanford, 2013). Dalam perhitungannya dapat dilakukan dengan menggunakan Persamaan 2.3.

$$P@K = \frac{|relevan \cap dokumen\ diperoleh\ sejumlah\ k|}{|dokumen\ diperoleh\ sejumlah\ k|} \quad (2.3)$$

2.7.4 *Recall*

Recall merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Perhitungan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi dapat diperoleh dengan menggunakan Persamaan 2.4.

$$R = \frac{|relevan \cap dokumen\ diperoleh|}{|relevan|} \quad (2.4)$$

2.7.5 *F-Measure*

Merupakan pengukuran yang menilai timbal balik antara *precision* dan *recall.* Perhitungannya dapat diperoleh dengan menggunakan Persamaan 2.5.

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{P}+\frac{1}{R}\right)} = \frac{2XPXR}{P+R} \quad (2.5)$$

Keterangan:

P : *precision*

R : r*ecall*

Precision merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem, sedangkan *recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

2.7.6 *Accuracy*

Merupakan metode evaluasi dari sebuah sistem dengan menghitung probabilitas pengkategorian jawaban (sesuai atau tidak sesuai) dengan seluruh dokumen. Penjelasan dari probabilitas tersebut dijelaskan pada Tabel 2.4.

Tabel 2.4 Pengelompokan evaluasi sistem

	Sesuai	Tidak Sesuai
Diambil	A	B
Tidak Diambil	C	D

Sedangkan untuk menghitung nilai *accuracy* dapat digunakan Persamaan 2.6.

$$Accuracy = \frac{(A+D)}{(A+B+C+D)} \quad (2.6)$$

2.7.7 *Manhattan Distance*

Terdapat beberapa metode yang dapat digunakan untuk menghitung jarak antara dua titik salah satunya adalah *Manhattan Distance.* Pada penelitian yang dilakukan oleh Gautama dibandingkan perhitungan jarak antara metode *Manhattan* dengan *Euclidean*, hasilnya diperoleh kesimpulan bahwa perhitungan menggunakan metode *Manhattan* memiliki waktu eksekusi lebih singkat dibandingkan dengan *Euclidean* (Gautama, et al., 2015). Sehingga pada penelitian ini digunakan metode *Manhattan* sebagai metode perhitungan jarak antara kandidat jawaban dengan kata kunci pertanyaan yang ada.

Manhattan Distance/ City Block Distance, merupakan salah satu teknik yang sering digunakan untuk menentukan kesamaan antara dua buah obyek. Pengukuran ini dihasilkan berdasarkan penjumlahan jarak selisih antara dua buah obyek dan hasil yang didapatkan dari *Manhattan Distance* bernilai mutlak. *Manhattan Distance* melakukan perhitungan jarak dengan cara tegak lurus (Gautama, et al., 2015).

Manhattan Distance merupakan persamaan yang digunakan untuk menghitung jarak diantara dua titik dengan cara mengambil nilai absolut dari jarak kedua koordinat. Perhitungan jarak menggunakan *Manhattan Distance* dapat diperoleh dengan Persamaan 2.7.

$$Distance = |a - c| + |b - c| \quad (2.7)$$

Keterangan:
a: Koordinat titik a
b: Koordinat titik b
c: Koordinat titik c