Pembuatan Tool Anotasi Kata Ganti Bahasa Arab Menggunakan Coreference Resolution

Rendy Andrian Saputra[1], Moch Arif Bijaksana[2], Donni Richasdy[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]rendyandrian@students.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id,
[3]donnir@telkomuniversity.ac.id

Abstrak

Al-Quran merupakan kitab suci orang muslim yang didalamnya banyak sekali ilmu pengetahuan. Seperti yang kita ketahui bahwa Al-Quran diturunkan dengan bahasa arab, sedangkan kita menggunakan bahasa Indonesia. Inilah salah satu penyebab yang membuat kebanyakan orang menjadi sulit memahami isi kandungan dalam Al-Quran. Mengetahui kesetaraan kata dari sebuah kata ganti sangat penting untuk memahami Al-Quran. Untuk mengetahui kesetaraan kata dari sebuah kata, diperlukannya Coreference Resolution. Coreference Resolution merupakan subtugas dari Natural Language Processing (NLP) yang bertugas untuk mengidentifikasi kesetaraan antar entitas, dengan menggunakan metode Naive Bayes sebagai metode klasifikasi yang telah terdapat di dalam tool anotasi. Tool anotasi diperuntukan untuk pengguna yang ahli dalam menafsirkan makna yang terkandung di dalam Al-Quran. Dimaksudkan agar hasil pemberikan rujukan disetiap kata memiliki data yang valid. serta dengan menggunakan tool anotasi pengguna dapat memberikan kesetaraan kata dari suatu kata ganti yang dapat mengacu pada suatu objek dikalimat sebelumnya. Berdasarkan hasil pengujian telah didapatkan nilai akurasi sebesar 80%.

Kata kunci : Naive Bayes, anotasi, tool,

Abstract

Al-Quran is a Muslim holy book in which a lot of knowledge. As we know that the Koran was revealed in Arabic, while we use Indonesian. This is one of the causes that makes it difficult for most people to understand the contents of the Koran. Knowing the equivalence of words from pronouns is very important to understand the Koran. To find out the equivalence of words from a word, Core Resolution is needed. Coreference Resolution is a Natural Language Processing (NLP) sub-task whose task is to identify equality between entities, using the Naive Bayes method as a classification method that is already contained in the annotation tool. Annotation tools are intended for users who are experts in interpreting the meanings contained in the Koran. Intended so that the results of giving references in each word have valid data. and by using annotation tools, users can provide pronoun equations that can refer to objects in the previous sentence. Based on the test results, an accuracy value of 80%

Keywords: Naive Bayes, anotasi, tool

1. Pendahuluan

Latar Belakang

Al-Quran merupakan kitab suci orang muslim yang didalamnya banyak sekali ilmu pengetahuan. Seperti yang kita ketahui bahwa Al-Quran diturunkan dengan bahasa arab, sedangkan kita menggunakan bahasa Indonesia. Inilah salah satu penyebab yang membuat kebanyakan orang menjadi sulit memahami isi kandungan dalam Al-Quran. Penerjemah Al-Quran mengalihkan pesan dari Al-Quran ke bahasa asing selain bahasa Arab agar dapat dipahami maksud dari firman Allah tersebut. Namun, banyak orang yang mengeluhkan bahwa hasil terjemahan sulit dipahami, terjemahan pada Al-Quran banyak sekali mengulang kata-kata yang sama pada suatu kalimat yang dimana membuat sulit mengartikan kata tersebut merujuk kemana. Misalkan terdapat kalimat :

Al-Baqarah: 21. Hai **manusia**, sembahlah Tuhan**mu** yang telah menciptakan**mu** dan orang-orang yang sebelum**mu**, agar **kamu** bertakwa.

Pada kalimat diatas, kata "**mu**" dan "**kamu**" merujuk ke kata "**manusia**". Untuk membantu pengguna yang ahli dalam memberikan padanan kata di dalam bahasa arab itu perlu menggunakan tool yang memungkinkannya

cepat mengidentifikasi entitas dalam kata ganti dan hubungan di antara kata sebelumnya. Sehingga dapat meminimalkan waktu yang diperlukan untuk memberikan padanan kata disetiap surat yang ada di dalam Al-Quran. [10] Proses ini dilakukan untuk menemukan apakah ungkapan tersebut merujuk ke entitas yang sama. Namun pada penelitian ini difokuskan untuk kata ganti pronoun. Maka dari itu dibutuhkan juga banyaknya kata ganti pronoun didalam Al-Quran. Semakin banyak kata dengan susunan kata disetiap surat maka sistem akan mampu mengetahui arti dari kata tersebut .Setelah memberikan padanan kata ganti pronoun lalu dilakukan Coreference Resolution dengan metode Naive Bayes. Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah dalam penelitian ini adalah bagaimana membuat tool untuk anotasi kata ganti pronoun pada Al-Quran dengan *Coreference Resolution*

Tujuan dari penelitian ini yaitu membuat tool untuk anotasi kata ganti pronoun pada Al-Quran dengan *Coreference Resolution* agar mempermudah pengguna untuk memahami Al-Quran.

Topik dan Batasannya

Berdasarkan dari latar belakang yang telah dijelaskan, topik-topik yang di angkat dalam tugas akhir ini sebagai berikut:

1. **Coreference Resolution**

   Hasil dari tugas akhir ini berupa hasil coreference resolution dari kata ganti berjenis pronoun. dimana kata ganti tersebut dapat merepresentasikan relasi antara istilah dengan objek yang lain. adapun contoh seperti berikut kata ayat 1:5:1 (إِيَّاكَ) artinya Hanya kepada Engkaulah yang merujuk dengan kata pada ayat 1:1:2 (اللَّـهِ) artinya Allah SWT. yang dimaksudkan bahwa *Hanya Kepada Engkaulah* memiliki arti sama dengan *Allah SWT.*

2. **Input dan Output**

   Inputan dari sistem ini berupa kata-kata yang ingin diperediksi kedekatan arti dengan kata-kata yang lain. Dan outputan dari sistem ini berupa kesamaan arti dari kata kata lain. Adapun contohnya seperti berikut ini: ayat 45:4:9 sebagai inputan 1 لِلْمُؤْمِنِينَ (bagi orang-orang yang beriman), inputan 2 kata pada 45:4:9 يُوقِنُونَ kemudian akan menghasilkan outputan nilai similarity yang tinggi karena kata-kata tersebut memiliki keterkaitan semantik yaitu sama-sama menjelaskan tentang kaum atau orang yang beriman atau orang-orang yang meyakini, begitu juga dengan kata-kata yang lainnya. Pada tabel terdapat beberapa contoh dari input dan output segaligus sebagai test set dari sistem yang akan dibangun. Dan untuk lebih lengkapnya dapat dilihat pada tabel yang terlampir pada lampiran 1.

Tabel 1. Contoh Input dan Output

No.	Kata (Input)	Rujukan (input)	Json/Text (Output)
1.	(إِيَّاكَ)	اللَّـهِ	{ ”Kata”:”(إِيَّاكَ), ”Terjemahan”:”You Alone, ”Rujukan”:اللَّـهِ }

Berdasarkan tabel diatas menjelaskan bahwa pengguna, dapat menginputkan kata yang akan dirujuk. Kemudian pengguna dapat memberikan persamaan dari kata tersebut. sehingga kata yang telah diberikan rujukan. Akan menampilkan hasil keluaran data berupa data JSON.

Adapun batasan dari permasalahan yang ada pada tugas akhir ini sebagai berikut:

1. Hanya menggunakan kata ganti berjenis pronoun.
2. Tidak semua kata ganti memiliki makna yang merujuk pada ayat tertentu.
3. Pengguna hanya dapat melabelkan kata ganti berjenis pronoun.
4. Tool ini tidak dapat mendeteksi kebenaran arti rujukan kata, dikarenakan pengguna dapat melabelkan secara mandiri.

Tujuan

Berikut adalah tujuan yang ingin dicapai pada penulisan proposal/TA.

1. Membuat tool untuk anotasi kata ganti pronoun pada Al-Quran dengan Coreference Resolution.
2. Dapat menentukan anteseden dari sebuah anaphor.

3. Pengguna dapat menentukan sendiri rujukan dari kata ganti berjenis pronoun.

Organisasi Tulisan

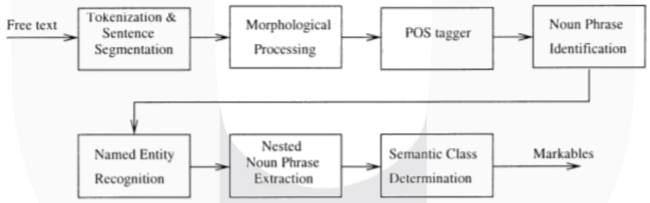
Pada sub-bagian ini dituliskan bagian-bagian selanjutnya (setelah Pendahuluan) pada jurnal TA ini, disertai penjelasan sangat singkat.

2. Studi Terkait

2.1 Ringkasan Studi Terkait

Jurnal ini di publikasi kan oleh Natalia N. Modjeska, Katja Markert, Malvina Nissim dengan judul Using the Web in Machine Learning for Other-Anaphora Resolution. Dimana jurnal tersebut menjelaskan tentang penyajian pendekatan pembelajaran mesin ke anafora lain, menggunakan classifier Naive Bayes (NB) dengan dua set fitur yang berbeda. yaitu dengan menggunakan website dan tidak dengan menggunakan website. Jurnal tersebut menggunakan classifier Naive Bayes, khususnya implementasi di perpustakaan Weka ML Data pelatihan dihasilkan mengikuti prosedur yang digunakan oleh Soon et al. (2001) untuk resolusi coreference. Setiap pasangan kata yang ingin dirujuk dan pendahulunya yang terdekat sebelumnya menciptakan contoh pelatihan positif. Untuk menghasilkan contoh pelatihan negatif, kami memasangkan anafor dengan masing - masing NP yang mengintervensi antara kata yang ingin dirujuk dan antesedennya(Kata yang serupa). Prosedur ini menghasilkan satu set 3.084 pasangan anteseden-anaforis, di antaranya 500 (16 contoh pelatihan positif). Pengklasifikasi dilatih dan diuji menggunakan 10 kali lipat validasi silang. Kami mengikuti praktik umum Algoritma ML untuk resolusi coreference dan menghitung presisi (P), recall (R), dan F-ukur (F) aktif.

Menurut jurnal yang berjudul A Machine Learning Approach to Coreference Resolution of Noun Phrases yang dibuat oleh Wee Meng Soon, Hwee Tou Ng, Daniel Chung Yong Lim. Yang mengadopsi pendekatan berbasis mesin, pembelajaran mesin untuk inti kata benda resolusi. Pendekatan ini membutuhkan kumpulan dokumen pelatihan yang relatif kecil telah dianotasi dengan rantai inti dari frase nomina. Semua kemungkinan yang dapat ditandai dalam dokumen pelatihan ditentukan oleh modul pemrosesan bahasa, dan contoh pelatihan dalam bentuk vektor fitur dihasilkan untuk pasangan yang sesuai dari kata yang dirujuk. Contoh-contoh pelatihan ini kemudian diberikan untuk membangun algoritma pembelajaran sebuah classifier. Untuk menentukan rantai coreference dalam dokumen baru, semua yang dapat ditandai adalah pasangan-pasangan yang ditentukan dan potensial dari tanda-tanda kata yang disajikan ke pengklasifikasi, yang memutuskan apakah kedua pasangan kata itu sebenarnya coreference



Gambar 1. Sistem Arsitektur Natural Language Processing (NLP)

2.2 Pre Prosessing

Pre Prosessing merupakan tahapan awal untuk melakukan proses pengolahan data asli sebelum melakukan tahapan utama dari Metode Latent Semantic Analysis (LSA). Tujuan adanya Pre Prosessing terbagi menjadi beberapa bagian diantaranya :

- **Pembersihan Data**
  1. Mengidentifikasi dan menghapus outlier
  2. Data noise yang halus
- **Transformasi Data**
  1. Normalisasi dan agregasi

di dalam proses pre prosessing terdapat tahapan yang dapat kita lakukan diantaranya : *Tokenization*

2.2.1 Tokenization

Tokenization atau biasa disebut dengan parsing, dimana proses tokenization merupakan suatu tahapan pemotongan kata berdasarkan kata yang menyusunnya. Pemisah antar kata dari proses tokenization adalah sebuah space (spasi)

Tabel 2. Contoh Tokenization

Data Input	Hasil Tokenization
Siapa yang paling bertakwa ?	(siapa)—(yang)—(paling)—(bertakwa)

2.2.2 Selection

Selection merupakan feature yang dapat digunakan untuk mengurangi atribut yang kurang relevan pada dataset. Beberapa algoritma feature selection yang digunakan adalah information gain, chi square, forward selection dan backward elimination.

- **Part Of Speech (POS) Tagging**

  Part Of Speech merupakan tahap pemberian suatu label dimana setiap kata akan diberikan label label sesuai dengan jenis kata. terdapat 8 jenis part of speech, diantaranya

  1. noun (kata benda)
  2. pronoun (kata ganti)
  3. verb (kata kerja)
  4. adjective (kata sifat)
  5. adverb (kata keterangan)
  6. preposition (kata depan)
  7. conjunction (kata hubung)
  8. int

2.2.3 Classification

Classification adalah sebuah metode dari data mining yang digunakan untuk memprediksi kategori atau kelas dari suatu data. Teks dianalisis oleh suatu model dan kemudian tag yang sesuai diterapkan berdasarkan konten. Model pembelajaran mesin yang dapat secara otomatis menerapkan tag untuk klasifikasi dikenal sebagai pengklasifikasi.

1. Naive Bayes

   Naive Bayes merupakan suatu metoda pembelajaran mesin yang digunakan untuk perhitungan probabilitas dan statistik. adapun kekurangan dan kelebihan menggunakan naive bayes

   - Kelebihan
     (a) Menangani kuantitatif dan data diskrit
     (b) Cepat dan efisiensi ruang
     (c) Bisa digunakan walaup atribut tidak relvan
   - Kekurangan
     (a) Mengasumsikan variabel bebas
     (b) Tidak berlaku jika probabilitas kondisionalnya adalah nol, apabila nol maka probabilitas prediksi akan bernilai nol juga

Teorema Naive Bayes

Teorema Bayes yang menjadi dasar dari metoda tersebut. Pada Teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan X dan H), maka Teorema Bayes dirumuskan sebagai berikut

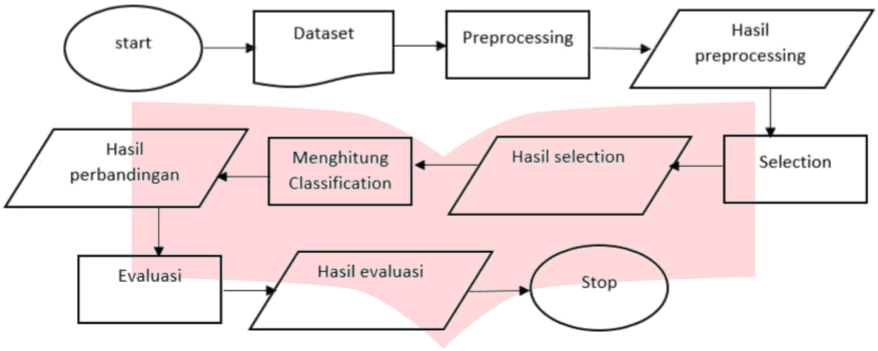
Dalam notasi ini P(H — X) berarti peluang kejadian H bila X terjadi dan P(X—H) peluang kejadian X bila H terjadi. Penerapan yang diharapkan dari rumus naive bayes ini adalah mencari probabilitas kata dengan mengkategorikan kata tersebut berdasarkan surat. Serta hasil akhir dari pencarian tersebut adalah mengkategorikan bahwa kata tersebut merupakan kata valid atau tidak valid. Berdasarkan anotasi kata pengguna yang dilakukan secara mandiri.

$$P(H|X) = \frac{P(X|H)}{P(X)}.P(H)$$

Gambar 2. Rumus Naive Bayes

3. Sistem yang Dibangun

Sistem yang dibangun bertujuan untuk menghasilkan tool yang dimana user dapat menentukan rujukan kata (PRONOUN) secara mandiri, serta hasil output dari tool tersebut berupa data yang bertipe JSON.



Gambar 3. Gambaran umum sistem

Berdasarkan gambar diatas menjelaskan bahwa data yang telah didapatkan akan melalui proses prosessing, dimana proses prosessing ini melewati tahapan tokenisasi, yaitu suatu tahapan pemisahan kalimat berdasarkan kata. Kemudian hasil pemisahan perkata tersebut akan melalui tahapan selection, Salah satu tahapan selection yang digunakan adalah Tokenisasi, dimana tahapan tersebut berguna untuk memberikan label disetiap kata yang telah terpisah. Sehingga di setiap kata akan memiliki label yang berbeda. Setelah melalui tahapan pemberian label. Maka kata tersebut akan melalui tahapan clasifikasi, yaitu perhitungan dengan metode naive bayes. Yang berguna untuk memprediksi probabilitas dari setiap kata di dalam Al-Quran. Hasil prediksi akan dihitung pada tahap evaluasi.

3.1 Dataset

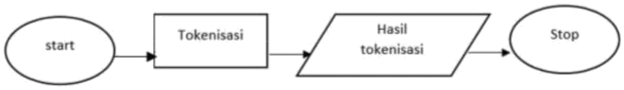
Data set yang digunakan pada tugas akhir ini didapat dari situs online yaitu www.corpus.quran.com dan ensiklopedia Al-Quran[11]. Dataset yang berisikan kata atau lemma yang ingin diketahui kedekatan makna dengan kata lainnya dan akan digunakan sebagai input dari sistem. Dan untuk data test dan data training di buat secara manual dari berbagai sumber online, ensiklopedia dan para ahlinya.[6]

3.2 Prepocessing

Pada tahap preprocessing akan dilakukan pemisahan lemma dari corpus yang didapat kemudian akan di lakukan case folding, tokenisasi pada lemma tersebut. Preprocessing dilakukan dengan tujuan memastikan dataset siap untuk diproses. [13]

3.2.1 Tokenisasi

berikut alur dalam melakukan tahap Tokenisasi (memisahkan antar kata berdasarkan kata)[15]



Gambar 4. Proses Tokenisasi

Berdasarkan gambar di atas menjelaskan bahwa data yang telah dimiliki akan melewati proses tokenisasi. Dimana proses tokenisasi merupakan proses pemisahan kalimat menjadi perkata.

Tabel 3. Contoh Tokenisasi

Sebelum Tokenisasi	Sesudah Tokenisasi
الرَّحْمَنالرَّحِيمِ	الرَّحِيمِ — الرَّحْمَن
maha pengasih maha penyanyang	maha pengasih — maha penyanyang

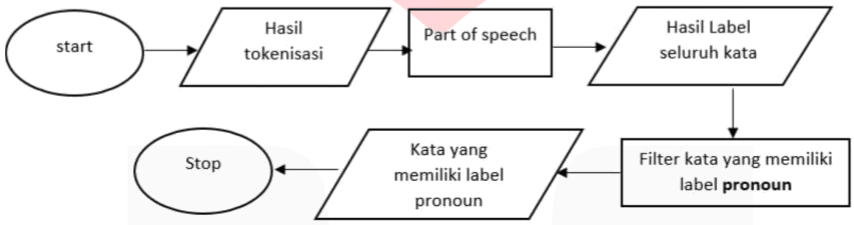
Dari hasil tabel diatas akan terlihat bahwasannya kalimat sesudah tertokenisasi akan terpisah menjadi perkata. Kemudian kata yang telah akan diberikan label melalui tahapan selection.

3.3 Selection

Pada tahap selection akan dilakukan pengurangan atribut yang tidak diperlukan. pada tahap selection dapat menggunakan Part Of Speech(POS) Tagging untuk mendapatkan suatu kata yang berlabel **pronoun**.[14]

3.3.1 Part Of Speech (POS) Tagging

Pada tahap Part Of Speech Tagging, kata yang sudah di tokenisasi akan diberikan suatu label. dimana label yang akan kita gunakan adalah kata yang berlabel **pronoun**. berikut alur penggunaan Part Of Speech Tagging. [5]



Gambar 5. Proses Part Of Speech

Berdasarkan gambar diatas menjelaskan bahwa hasil dari proses tokenisasi akan melalui tahapan Part Of Speech (POS) Tagging. Dimana tahapan ini akan memberikan suatu label disetiap kata yang telah tertokenisasi. Sehingga kata akan memiliki label yang berbeda-beda sesuai dengan kata ganti dari kata tersebut.

Tabel 4. Contoh Part Of Speech

Sebelum Part Of Speech	Sesudah Part Of Speech
إِيَّاكَ	إِيَّاكَ—PROUN
نَعْبُدُ	نَعْبُدُ—V
وَإِيَّاكَ	وَإِيَّاكَ—PROUN CONJ
نَسْتَعِينُ	نَسْتَعِينُ—V

dari proses Part Of Speech ini, kita dapat memberikan teknik Coreference Resolution,[12] untuk mendapatkan sumber rujukan secara mandiri dari kata yang berlabel pronoun.

Tabel 5. Contoh Coreference Resolution [9]

Urutan	Kata	Terjemahan	Kata Ganti	Rujukan
(1:2:2)	اللَّهِ	Allah	PN	اللَّهِ
(1:5:1)	إِيَّاكَ	Hanya kepada **Engkaulah**	PROUN	اللَّهِ
(1:5:3)	وَإِيَّاكَ	dan hanya kepada **Engkaulah**	PROUN CONJ	اللَّهِ

Gambar 6. Proses Coreference Resolution

3.4 Classification

di tahapan classification kita dapat menentukan probabilitas terkait pemahaman seseorang untuk memberikan suatu rujukan kata prounoun di alquran, menggunakan Naive Bayes. [1][3]

3.4.1 Naive Bayes

Naive Bayes merupakan suatu metode untuk menentukan probabilitas dan statistik. Berikut alur penerapan metode Naive Bayes untuk menentukan probabilitas.[7]

Tabel 6. Contoh Data Training

Urutan	Kata	Terjemahan	Rujukan	Validasi
(1:2:2)	اللَّهِ	Allah	اللَّهِ	Valid
(1:5:1)	إِيَّاكَ	Hanya kepada **Engkaulah**	اللَّهِ	Valid

Tabel 7. Contoh Data Testing

Urutan	Kata	Terjemahan	Rujukan	Validasi	Kata Ganti
(1:2:2)	أَنْعَمْتَ	telah **Engkau** beri nikmat	-	Tidak Valid	PRON
(1:5:3)	وإياك	dan hanya kepada **Engkaulah**	اللَّهِ	Valid	PRON CONJ
(1:6:1)	اهدنا	Tunjukilah **kami**	أَنْعَمْتَ	Valid	PRON V
(92:11:3)	عَنْهُ	dari padanya/bagi**nya**	مَن	Valid	PRON P
(92:11:4)	مَالُهُ	dan hanya kepada **Engkaulah**	-	Tidak Valid	PRON N
(92:13:40)	لَنَا	maka aku peringatkan kamu	-	Tidak Valid	PRON P
(95:4:2)	خَلَقْنَا	**kami** telah menciptakan	الله	Valid	PRON V
(95:6:4)	وَعَمِلُوا	dan **mereka** berbuat/beramal	الَذِيْنَ	Valid	PRON V
(95:6:6)	فَلَهُمْ	maka bagi **mereka**	الَذِيْنَ	Valid	PRON P
(95:7:2)	يُكَذِبك	mendustakan**mu**	الإِنْسَانَ	Valid	PRON V

Terdapat 2 validasi dari sebuah data testing, yaitu kata valid dan tidak valid. Dimana dimaksudkan kata yang dapat dikatakan valid, merupakan suatu kata yang memiliki suatu rujukan dengan kata lain berdasarkan masing masing surat yang terkandung di dalam Al-Quran. serta dengan perhitungan naive bayes, kita dapat memperoleh suatu nilai yang bisa dievaluasi terkait pemahaman seseorang terhadap pemberian rujukan kata ganti pronoun

P(Valid)	= 7/10	= 0.7
P(Tidak Valid)	= 3/10	= 0.3
P(Surat 1 | Valid)	= 2/7	= 0.28
P(Surat 1 | Tidak Valid)	= 1/3	= 0.33
Kemudian,		
P(Valid) x P(Surat 1 | Valid)	= 0.7 x 0.28	= 0.19
P(Tidak Valid) x P(Surat 1 | Tidak Valid)	= 0.3 x 0.33	= 0.09

P(Surat 92 | Valid)	= 1/7	= 0. 14
P(Surat 92 | Tidak Valid)	= 2/3	= 0. 66
Kemudian,		
P(Valid) x P(Surat 92 | Valid)	= 0.7 x 0. 14	= 0.09
P(Tidak Valid) x P(Surat 92 | Tidak Valid)	= 0.3 x 0. 66	= 0.198

P(Surat 95 | Valid)	= 4/7	= 0. 57
P(Surat 95 | Tidak Valid)	= 0/3	= 0
Kemudian,		
P(Valid) x P(Surat 95 | Valid)	= 0.7 x 0. 57	= 0.39
P(Tidak Valid) x P(Surat 95 | Tidak Valid)	= 0.3 x 0	= 0

4. Evaluasi

Precision dan **Recall** merupakan langkah yang banyak digunakan untuk mengevaluasi kualitas hasil. Precision digunakan untuk mengukur tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem sedangkan recall digunakan untuk mengukur tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. untuk mengukur keduanya maka dapat menggunakan akurasi. Dimana akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dan nilai aktual.[4][8]

secara umum precision, recall dan akurasi dapat dirumuskan sebagai berikut :

		Nilai Sebenarnya	
		True	False
Nilai Prediksi	TRUE	TP (True Positive)	FP (False Positive)
	FALSE	FN (False Negative)	TN (True Negative)

Gambar 7. Gambaran Umum Precision, Recall dan Akurasi

Rumus Precision, Recall dan Akurasi

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Adapun jurnal yang menggunakan Precision dan Recall untuk menentukan hasil evaluasi, salah satu nya jurnal yang berjudul **Machine Learning Approaches to Coreference Resolution**. Dimana di dalam jurnal tersebut menjelaskan bahwa Precision dan Recall merupakan langkah yang banyak digunakan untuk mengevaluasi kualitas hasil proses pemberian rujukan dari suatu kata [4]

4.1 Hasil Pengujian

didalam tugas akhir ini terdapat contoh data yang akan diujikan. data yang akan diujikan harus melalui tahap case folding, tokenisasi serta Part Of Speech Tagging. yang dimana terdapat suatu proses pemberian suatu label di setiap kata. Yang nantinya kata tersebut akan di evaluasi akurasi kebenarannya. Data yang akan diujikan sebagai berikut : [2]

Gambar 8. Proses Coreference Resolution

Urutan	Kata	Terjemahan	Rujukan	Validasi
(1:2:2)	اللَّهِ	Allah	اللَّهِ	Valid
(1:2:3)	اللَّرَبِ	Tuhan	اللَّهِ	Valid
(1:3:1)	الرَّحْمَنِ	Yang **Maha Pengasih**	اللَّهِ	Valid
(1:3:2)	الرَّحِيمِ	Yang **Maha Penyayang**	اللَّهِ	Valid
(1:5:1)	إِيَّاكَ	Hanya kepada **Engkaulah**	اللَّهِ	Valid
(1:5:3)	وَإِيَّاكَ	dan hanya kepada **Engkaulah**	اللَّهِ	Valid
(1:6:1)	اهْدِنَا	Tunjukilah **kami**	-	**Tidak Valid**
(1:7:2)	االَّذِينَ	**Orang Orang** yang	-	**Tidak Valid**
(1:7:4)	عَلَيْهِمْ	Kepada**nya**	-	**Tidak Valid**
(1:7:7)	عَلَيْهِمْ	**Kepada** mereka	-	**Tidak Valid**
(1:7:9)	الضَّالِّينَ	**Mereka** yang sesat	-	**Tidak Valid**

Studi kasus pengujian. terdapat 503 data training yang terdiri dari 58 data valid dan 445 data non valid. kemudian terdapat data tes yang telah dideskripsikan oleh user, yang dimana terdapat 4 kali percobaan untuk menentukan perbandingan antar masing masing user.

Percobaan 1	Jumlah Data	Valid	Tidak Valid
Data Training	503	58	445
Data Tes	302	30	272

Percobaan 2	Jumlah Data	Valid	Tidak Valid
Data Training	503	58	445
Data Tes	302	58	244

Percobaan 3	Jumlah Data	Valid	Tidak Valid
Data Training	503	58	445
Data Tes	202	30	172

Percobaan 4	Jumlah Data	Valid	Tidak Valid
Data Training	503	58	445
Data Tes	102	40	62

4.2 Analisis Hasil Pengujian

Dari hasil pengujian di atas, dapat kita jabarkan berdasarkan tabel nilai prediksi. Maka hasil yang diperoleh sebagai berikut :

Hasil Percobaan	Precision	Recall	Accuracy
Percobaan 1	9%	34%	59%
Percobaan 2	19%	50%	62%
Percobaan 3	14%	34%	67%
Percobaan 4	39%	40%	80%

Berdasarkan percobaan diatas maka dapat kita lihat perolehan percobaan ke 4 memiliki akurasi yaitu sebesar 80%. Perolehan nilai akurasi tertinggi ini dipengaruhi dari beberapa faktor, salah satunya adalah banyaknya ketersediaan data yang akan diolah. Semakin besar jumlah data Train dan semakin sedikit data yang dites maka akan semakin baik pula model yang dihasilkan, sehingga nilai akurasi yang dihasilkan juga akan meningkat.

5. Kesimpulan

Kesimpulan dari Tugas Akhir ini adalah pengguna sudah dapat menggunakan tool secara mandiri. Dimana pengguna aplikasi tool anotasi ini diperuntukkan kepada pengguna yang ahli dalam menafsirkan kata yang terkandung di dalam Al-Quran. Diharapkan kata yang teranotasi merupakan kata yang memiliki makna yang valid. Sehingga dapat memperoleh hasil yang baik dalam penganotasian disetiap kata. Kata yang dapat diberikan rujukan merupakan kata yang berjenis pronoun. penulis telah melakukan suatu percobaan yang dimana dapat menghasilkan akurasi sebesar 80% berdasarkan ketepatan pengguna untuk menganotasi kata, dengan menggunakan metode naive bayes sebagai metode klasifikasi.

Daftar Pustaka

[1] S. W. B. Chinatsu Aone. Evaluating automated and manual acquisition of anaphora resolution strategies. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995.

[2] L. G. Christian HARDMEIER1. A graphical pronoun analysis tool for the protest pronoun evaluation test suite. *Baltic J. Modern Computing*, 4:318–330, 2016.

[3] K. W. Claire Cardie. Noun phrase coreference as clustering. *Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[4] N. G. L. Machine learning approaches to coreference resolution. *Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics,Prague, Czech Republic.*, 2008.

[5] H. T. N. M. Soon and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[6] A. B. Muhammad. Annotation of conceptual co-reference and text mining the qur'an. *IEEE Intelligent Systems*, 2012.

[7] M. N. Natalia N. Modjeska, Katja Markert. Using the web in machine learning for other-anaphora resolution. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003.

[8] Z. Z. Nguy Giang Linh. Rule-based approach to pronominal anaphora resolution applied on the prague dependency treebank 2.0 data. *Proceedings of DAARC 2007*, 2007.

[9] M. Poesio. The mate/gnome scheme for anaphoric annotation. *in Proceedings of SIGDIAL*, 2004.

[10] C. O. C. B. L. J. Ruslan Mitkov, Richard Evans and V. Sotirova†. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*, 2000.

[11] Sahabuddin, M. Q. Shihab, and Sahabuddin. *Ensiklopedia Al-Qur'an: kajian kosakata*. Lentera Hati, 2007.

[12] L. B. Souha Hammami and A. B. Hamadou. Arabic anaphora resolution: Corpora annotation with corefe-rential links. *The International Arab Journal of Information Technology*, 2009.

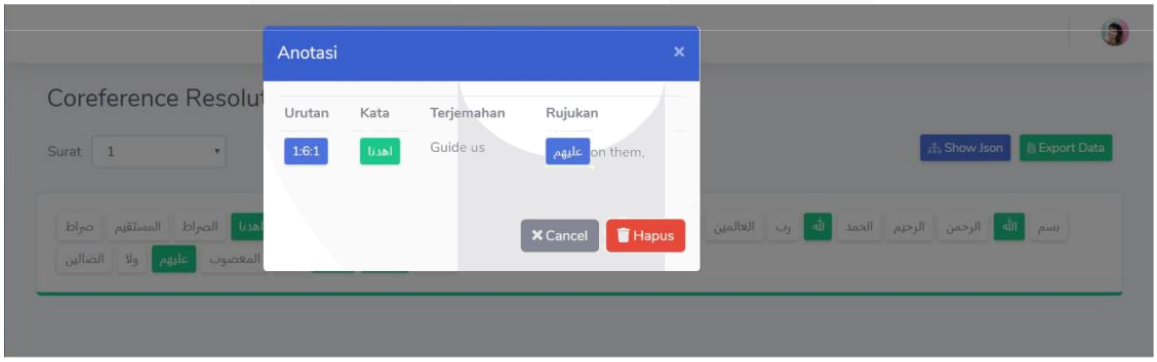
[13] J. udita Preiss. Anaphora resolution with word sense disambiguation. *Proceedings of SENSEVAL-2 Second International*, 2002.

[14] C. C. Vincent Ng. Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[15] D. C. Y. L. Wee Meng Soon, Hwee Tou Ng. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, page 521–544, 2001.

Lampiran

Gambar 9. Halaman Pilih Surat



Gambar 10. Hasil Kata yang Memiliki Rujukan