PENDEKATAN CENTERED SENTENCE-BASED CLUSTERING UNTUK SISTEM TANYA JAWAB PADA KASUS FRONT OFFICE

Zamaludin Abdulah[1], Ken Kinanti Purnamasari[2]
[1,2] Universitas Komputer Indonesia
Jalan Dipatiukur 114-116 Bandung 40132
E-mail : zamaludin@live.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRAK

Penelitian tentang sistem tanya jawab sudah dilakukan sebelumnya yang menggunakan metode Centered Sentence-Based Clustering. Pada penelitian tersebut menggunakan data bahasa Inggris. Hal ini juga menjadi salah satu dasar penelitian ini dilakukan. Selain itu juga sitem tanya jawab harus dapat menangani pertanyaan yang ada pada lingkungan *front office* serta menangani kasus *out of vocabulary* (OOV).

Pada penelitian ini sistem yang dibanugn memiliki dua tahapan utama, yaitu *Clustering* dan pencarian jawaban. *Clustering* adalah tahapan awal dimana data latih akan diproses sehingga menghasilkan aturan yang akan menjadi pembanding pada tahap pencarian jawaban. *Clustering* akan mengelompokan pertanyaan, untuk kemudian mencari jawaban yang ada didalam database berdasarkan kedekatan kalimat pertanyaan dengan yang ada pada database. Hasil dari main process ini adalah jawaban yang telah diisikan sebelumnya pada setiap data latih.

Hasil menunjukan akurasi yang didapat pada pendekatan clustering untuk sistem tanya jawab pada kasus front office mendapat nilai akurasi sebesar 81,08%. Nilai akurasi tersebut disebabkan karena kurangnya data latih pada proses *clustering.*

Kata kunci : tanya jawab, *clustering, centered-sentence based clustering, rule-based,* bahasa Indonesia

1. PENDAHULUAN

Pada penelitian sebelumnya yang dilakukan pengujian terhadap sistem tanya jawab dengan menggunakan metode *Rule-Based.* Hasilnya diperoleh bahwa sistem dapat menjawab 25,88% pertanyaan dengan tepat. Penelitian tersebut mencoba menggabungkan metode *word-matching* dan *rule-based*[2]. Masalah yang timbul dari metode tersebut adalah kurangnya aturan dan ketika tidak ada aturan yang cocok untuk pertanyaan maka jawaban tidak akan didapatkan meskipun maksud dari pertanyaannya mirip dengan yang ada pada *rule* atau bisa disebut bahwa penelitian tersebut belum bisa menangani kasus *Out of Vocabulary* (OOV).

Untuk mengatasi masalah OOV, beberapa penelitian telah menggunakan *clustering* untuk mendapatkan jawaban dari pertanyaan yang diberikan. Dengan memanfaatkan kedekatan antara pertanyaan, *clustering* dapat meminimalisir terjadinya OOV[3]. Seperti pada penelitian lain yang menggunakan pendekatan metode *clustering* yaitu *Centered Sentence-based (CESED)* pada sistem tanya jawab yang menggunakan metode *rule based*. Hasil yang didapat yakni nilai akurasi sebesar 84.62% pertanyaan bertipe "*WHERE-type questions*"[4].

2. ISI PENELITIAN

Bagian ini dapat dibagi dalam beberapa sub pokok pembahasan sesuai dengan kebutuhan tulisan.

2.1 Sistem Tanya Jawab

Sistem Tanya Jawab (QA) merupakan sebuah sistem pada komputer yang berisi sebuah cara otomatis untuk menjawab pertanyaan menggunakan bahasa alami yang biasa dipakai manusia. Jawaban ini berisi informasi yang berasal dari sebuah sumber basis data [1]. Sebuah sistem tanya jawab diharapkan dapat membantu penggunanya untuk mendapatkan informasi tanpa harus bertanya kepada orang lain atau orang yang terkait. Bahkan contoh produk yang sudah ada seperti Kata.ai, sistem tanya jawab dijadikan sebagai asisten pribadi bagi penggunanya yang dapat melakukan hal-hal yang umumnya bisa dilakukan oleh asisten pribasi seperti pembukuan tiket transportasi dan hotel[7]. Pada penelitian ini juga akan dibangun sebuah sistem tanya jawab untuk kasus di Front Office Unikom.

2.2 Clustering

Pada dasarnya clustering terhadap data adalah suatu proses untuk mengelompokkan sekumpulan data tanpa suatu atribut kelas yang telah didefinisikan sebelumnya, berdasarkan pada prinsip konseptual clustering yaitu memaksimalkan dan juga meminimalkan kemiripan intra kelas[10]. Misalnya, sekumpulan obyek-obyek komoditi pertama-tama dapat clustering menjadi sebuah himpunan kelas-kelas dan lalu menjadi sebuah himpunan aturan-aturan yang dapat diturunkan berdasarkan suatu klasifikasi tertentu. Pada penelitian ini clustering

digunakan sebagai proses utama dalam pembentukan aturan sistem tanya jawab.

2.3 Centered Sentence-based Clustering

Strategi learning dari CESED adalah untuk mempelajari aturan pada sebuah bagian pada contoh data latih yang diberikan.

Tabel 2. 1 Contoh data latih

Pair No	Main-Sentence	Co-Sentence	Nilai Cosine
	Question No.	Question No.	
1	1	2	0,73
2	1	3	0,721
3	1	4	0,757
4	2	3	0,95
5	2	4	0,803
6	3	4	0,816

Main-Sentence merupakan kalimat yang akan menjadi center cluster yang dipilih berdasarkan urutan kalimat pertanyaan dan co-sentence adalah kalimat kandidat yang nantinya akan dihitung apakah masuk kedalam cluster atau tidak.

Clustering dimulai dari menghitung kalimat pertanyaan dari data latih. Main-sentence akan menjadi center pada sebuah cluster. Dengan kata lain, tujuan dari proses clustering ini adalah untuk menilai Co-sentence memiliki kemiripan dengan Main-sentence dibandingkan dengan semua Co-sentence lainnya. Pada tujuan ini, nilai threshold telah ditentukan.

2.4 TFIDF

Metode TF-IDF merupakan metode untuk menghitung bobot kedekatan setiap kalimat pertanyaan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat[6]. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Metode ini akan menghitung bobot setiap token t di dokumen d dengan rumus:

$$idf = \log(\frac{N}{df})$$

D adalah jumlah dokumen yang berisi term (t) dan idf adalah jumalh kemunculan (frekuensi) term terhadap D. Adapun algoritma yang digunakan untuk menghitung bobot (w) masing-masing dokumen terhadap kata kunci (query) menggunakan persamaan berikut.

$$w_{d.t} = tf_{d.t} * IDF_t$$

Dimana :
D : dokumen ke-d
t : term ke-t dari kata kunci
tf : term frekuensi/ frekuensi kata
w : bobot dokumen ke-d terhadap term ke-t

2.2 *Cosine Similarity*

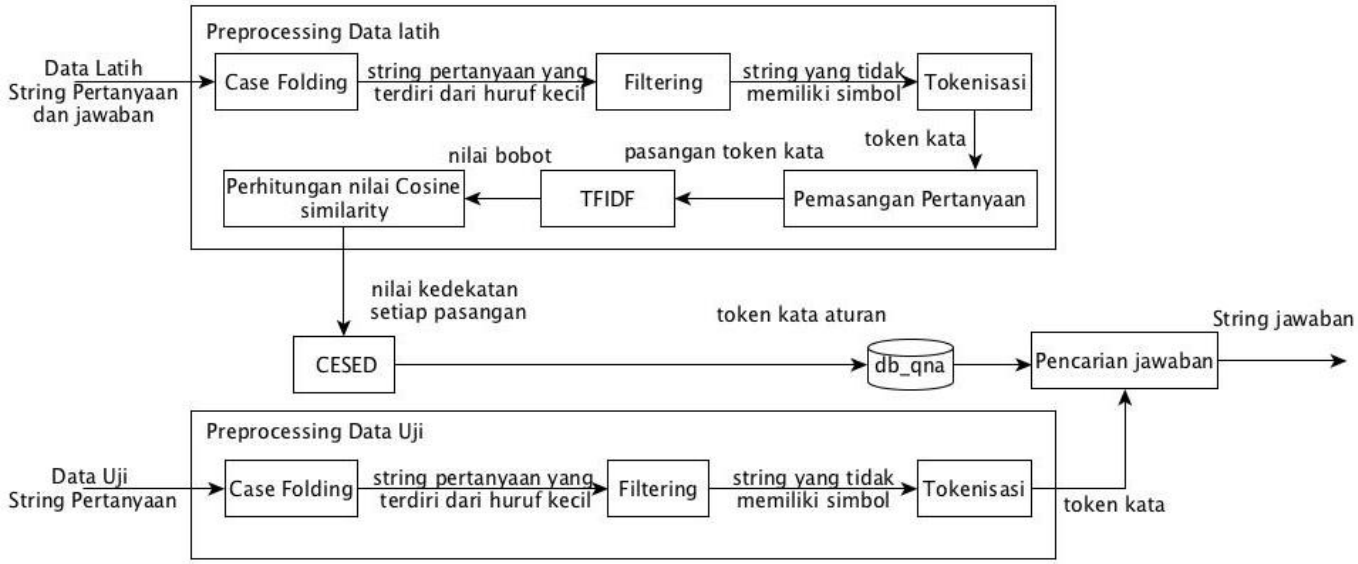
Cosine similarity digunakan untuk menghitung pendekatan relevansi *query* terhadap kalimat. Semakin besar nilai kesamaan vektor query dengan vektor dokumen maka *query* tersebut dipandang semakin relevan dengan dokumen. Cosinus dari 0^0 adalah 1, dan kurang dari 1 untuk setiap sudut lainnya. Dengan demikian dua vektor dengan orientasi yang sama memiliki kesamaan cosinus dari 1, dua vektor pada 90^0 memiliki kesamaan 0. Cosine similarity terutama digunakan dalam ruang positif, dimana hasilnya dibatasi (0,1). Cosine similarity kemudian memberi tolak ukur seberapa mirip dua dokumen [7].

Cosinus dari dua vektor dapat diturunkan dengan menggunakan dot product Euclidean yaitu 1.b = ||a|| ||b|| cos Θ dimana vektor atribut, A dan B kesamaan dan besarnya ditukan dengan persamaan (2.3).

$$CS(b_1, b_2) = \frac{\sum_{t=1}^{n} W_{t.b1} W_{t.b2}}{\sqrt{\sum_{t=1}^{n} W_{t.b1}^2} \sqrt{\sum_{t=1}^{n} W_{t.b2}^2}}$$

2.3 Analisis Masalah

Berdasarakan pendahuluah penelitian mengenai pendekatan centered sentence-based clustering untuk sistem tanya jawab pada kasus front office belum pernah dilakukan sebelumnya. Berdasarkan hal tersebut maka analisis proses digamabarkan pada gambar 2.1



Gambar 2.1 Skema Proses

2.4 Analisis Data Masukan

Data masukan yang akan diproses oleh sistem baik data latih maupun data uji diambil berdasarkan hasil dan wawancara. Adapuh format *file* yang akan dibaca yakni data berupa teks dan data tersebut dimasukan ke dalam *database*.

Tabel 2.2 Contoh Data Masukan Pertanyaan dan Jawaban.

NO	Pertanyaan	Jawaban
1	bagaimana cara mendaftar autodebet di unikom? :)	Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda
2	Bagaimana jika telat autodebet	Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda
3	bagaimana cara mendaftar Kuliah	Pendaftaran dapat dilakukan menggunakan perndaftaran online lalu menyerahkan formulir ke bagian PMB UNIKOM
4	kenapa Autodebet harus dibayar	Autodebet bertujuan untuk memudahkan pembayaran kuliah dimana mahasiswa cukup mentransfer sejumlah uang sesuai dengan biaya perkuliahan ke rekening mahasiswa tersebut

2.5 *Preprocessing*

Pada proses ini dijelaskan tentang penyaringan data dan ekstraksi data agar pada saat proses selanjutnya data sudah bersih atau ideal untuk diproses serta fitur yang ada pada data tersebut telah dapat digunakan. Gambaran umum tahap preprocessing dapat dilihat pada gambar 2.2.



Gambar 2. 2 Preprocessing

Pada penelitian ini *preprocessing* memiliki beberapa proses yakni :

- *Case folding* adalah proses penyetaraan jenis teks, pada penelitian ini huruf diubah menjadi huruf kecil (*lowercase*) agar tidak ada masalah dalam pencocokan kalimat pada proses selanjutnya.
- Proses *Filtering.* Pada proses ini data input yang berisi simbol dan *emoticon* akan dihilangkan karena tidak diperlukan dan dapat menghambat pada proses selanjutnya, dengan ketentuan proses penghilangan simbol akan menghilangkan karakter selain abjad a-z dan 0-9.

- Setelah melalui proses *filtering*, tahapan seanjutnya hasil dari proses *filtering* akan masuk ke proses tokenisasi. Proses tokenisasi pada penlitian ini adalah memisakah setiap token menggunakan delimeter spasi. Maka data yang akan dihasilkan dari tokenisasi adalah himpunan kata atau pecahan kata.

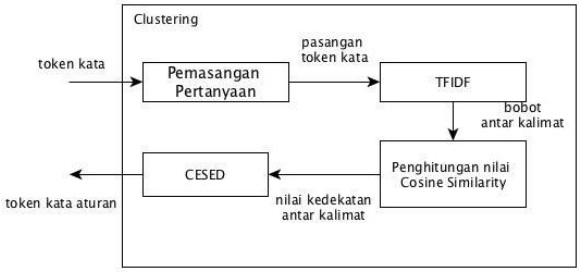
Setelah melewati preprocessing maka data yang dihasilkan adalah data token pada setiap data latih.

Tabel 2.3 Contoh Data Masukan Pertanyaan dan Jawaban.

i	P_i	Sebelum Tokenisasi	Setelah Tokenisasi
1	P_1	bagaimana cara mendaftar autodebet di unikom	bagaimana cara mendaftar autodebet di unikom
2	P_2	ada berapa kah fakultas di unikom	ada berapa kah fakultas di unikom
3	P_3	berapa biaya kuliah tahun 2016	berapa biaya kuliah tahun 2016
4	P_4	bagaimana cara mendaftar k/uliah	bagaimana cara mendaftar kuliah

2.6 *Clustering*

Proses *Clustering* yang akan dilakukan akan menggunakan metode CESED, pada penelitian ini *clustering* digunakan pada 2 proses yakni pembentukan aturan dan pada pencarian jawaban. Proses pembentukan aturan menggunakan clustering dapat dilihat pada gambar 2.4.



Gambar 2. 4 Skema Clustering

Untuk menentukan Main Class dan Kandidat Class maka proses pemasangan pertanyaan dilkaukan. Berikut adalah hasil pemasangan pertanyaan. Pada CESED Main Class disebut semagai Main-sentence dan Kandidat Class sebagai Co-Sentence.

Tabel 2.4 Hasil Pemasangan pertanyaan

Pasangan No	*Main-Sentence*	*Co-Sentence*
1	P_1	P_2
2	P_1	P_3
3	P_1	P_4
4	P_2	P_3
5	P_2	P_4
6	P_3	P_4

Untuk melakukan pembobotan kata maka proses TF.IDF dilakukan. Hasil penghitungan TF.IDF dapat dilihat pata tabel berikut. Pada contoh TF.IDF digunakan contoh pasangan no 3.

Berikut adalah tahapan penghitungan bobot TF.IDF.

1. Untuk mengisi kolom tf, hitung kemunculan kata "autodebet" pada setiap kalimat. Kata "autodebet" hanya muncul pada kalimat 1 saja, sehingga pada tf-P_1 bernilai 1 dan tf.S lainnya diberi nilai 0.
2. Selanjutnya kolom df, jumlahkan kemunculan kata "autodebet" dalam pasang pertanyaan. Karena kata "autodebet" hanya muncul 1 kali pada kalimat 1, maka df bernilai 1.
3. Mengisi kolom idf, yaitu dengan rumus (2.1) dengan memasukkan nilai n (jumlah kata) dan df (proses no.2). Nilai $n = 7$, dan $df = 2$, sehingga idf kata "autodebet" $= log(7/2) = 0.35218$
4. Mengisi kolom w, yaitu dengan rumus (2.2) dengan memasukkan nilai tf (proses no.1) dan idf (proses no.3). Sehingga mengisi w-$P_1 = tf$-$P_1.idf = 1.1{,}89209 = 1{,}89209$ dan w-P selanjutnya penghitungan $TF.IDF$ dapat dilihat pada Tabel

Tabel 2.5 Hasil TFIDF

Term	W = TF.IDF		W_{P1}^2	W_{P4}^2	$W_{P1}.W_{P4}$
	P_1	P_4			
bagaimana	0,544068	0,544068	0,29601	0,29601	0,29601
mendaftar	0,544068	0,544068	0,29601	0,29601	0,29601
cara	0,544068	0,544068	0,29601	0,29601	0,29601
autodebet	0,845098	0	0,714191	0	0
di	0,845098	0	0,714191	0	0
kuliah	0	0,544068	0	0,845098	0
unikom	0,845098	0	0,714191	0	0
$\sum$			3,0306	1,60222	0,88803

Setelah mendapatkan nilai bobot dari proses TF.IDF langkah selanjutnya adalah menghitung nilai kedekatan dengan menggunakan metode Cosine Similarity.

Tabel 2.5 Hasil Pencarian W^2

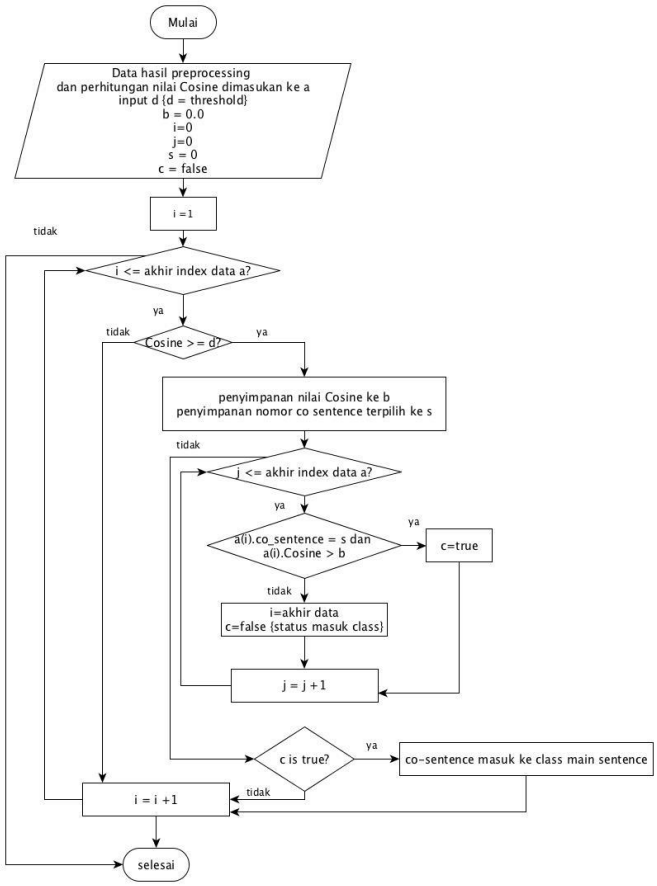
Term	TF			IDF	W= TF.IDF	
	P1	P4	df	Log(n/df)	P1	P4
bagaimana	1	1	2	0,544068	0,544068	0,544068
mendaftar	1	1	2	0,544068	0,544068	0,544068
cara	1	1	2	0,544068	0,544068	0,544068
autodebet	1	0	1	0,845098	0,845098	0
di	1	0	1	0,845098	0,845098	0
kuliah	0	1	1	0,845098	0	0,544068
unikom	1	0	1	0,845098	0,845098	0

Berikut adalah hasil penghitungan cosine similarity pada setiap pasangan.

Tabel 2. 6 Hasil Cosine Similarity

i	*Main-Sentence*	*Co-Sentence*	*Sim(Main-Sentence, Co-Sentence)*
1	P_1	P_2	0
2	P_1	P_3	0,40299
3	P_1	P_4	0,790021
4	P_2	P_3	0,370568
5	P_2	P_4	0
6	P_3	P_4	0

Setalah nilai kedekatan didapatkan langkah selanjut nya adalah melakukan cluster menggunakan metode CESED. Untuk langkah kerja CESED dapat dilihatp pada flowchar berikut.



Gambar 2.6 Flowchart CESED

Langkah awal pada proses CESED adalah dengan membandingkan setiap nilai *Cosine Similarity* pada tabel 3.8 lalu membandingkan dengan nilai *threshold* yang telah ditentukan yakni 0.6. Dimulai dari pasangan satu apakah nilai *Cosine Similarity* ≥ *threshold* atau tidak. Pada tabel dapat dilihat nilai *Cosine Similarity* pada setiap pasang kalimat berserta pasangan *Main-sentence* dan *Co-sentence*. Karena main-Sentence diukur sebagai center class yang potensial maka main Sentence akan dianggap sebagai cluster dan pada tahapan ini class dimasukan kedalam variabel C_i. Setiap cluster memiliki anggota yakni main-sentence itu sendiri.

Tabel 2. 8 Pembagian Cluster

i	*Cluster*	*Co-Sentence*	*Sim(Main-Sentence, Co-Sentence)*
1	C_1	P_2	0
2	C_1	P_3	0,40299
3	C_1	P_4	0,790021
4	C_2	P_3	0,370568
5	C_2	P_4	0
6	C_3	P_4	0

Setelah diketahui nilai Cosine Pasangan 1 maka dibandingkan apakah nilai *Cosine Similarity* tersebut ≥ 0.6. Karena nilai Sim < 0.6 maka lanjut ke nilai Cosine pasangan selanjutnya. Karena nilai COSINE < *threshold* maka lanjut ke nilai Cosine pasangan selanjutnya. Karena nilai COSINE ≥ *threshold* maka dicek apakah nilai COSINE tersebut adalah nilai terbesar dari semua dibandingkan nilai COSINE lain pada *Co-Sentence* yang sama. Karena nilai COSINE pasangan 5 ≤ nilai COSINE pasangan 3 maka perbandingan dilakukan kembali pada pasangan yang memiliki nilai *co-sentence* = 4. Karena nilai COSINE pasangan 6 ≤ nilai COSINE pasangan 3 maka dan tidak ada lagi pasangan yang memiliki nilai *co-Sentence* 4 maka P_4 berkerumun dengan grup P_1. Setelah itu pasangan yang memiliki nilai *co-sentence* = 4 tidak akan bandingkan kembali karena telah masuk ke *cluster* pertanyaan 1 dilanjut dengan menghitung nilai COSINE berikutnya.

Tabel 2. 9 Hasil Proses Clustering

Cluster	Co Sentence
C_1	P_1
	P_4
C_2	P_2
C_3	P_3

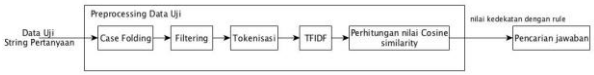
Dari hasil *clustering* tersebut menjadi aturan baru bagi sistem tanya jawab diambil dari pertanyaan yang tersisa setelah melewati clustering, data pertanyaan yang tersisa yakni P_1, P_2 dan P_3. Maka aturan yang akan muncul seperti pada tabel 2.10

Tabel 2. 10 Aturan yang Muncul Setelah Clustering

i	Pertanyaan	Jawaban	R_i
1	bagaimana cara mendaftar autodebet di unikom	Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda	R_1
2	ada berapa kah fakultas di unikom	Unikom memiliki 6 Fakultas dan 24 Program Studi	R_2
3	berapa biaya kuliah tahun 2016	Untuk Semester ganjil biaya kuliah sebesar 5.300.000 untuk fakultas teknik	R_3

2.6 Pencarian Jawaban

Pada proses ini dulakukan kembali proses *Preprocessing* tetapi terhadap data uji. Setelah itu hasil dari preprocessing akan masuk ke proses pencarian jawaban dan akan menghasilkan jawaban dari sistem.



Gambar 2. 7 Pencarian Jawaban

Berikut adalah contoh data uji yang telah melwati *preprocessing.* Dapat dilihat pada tabel 2.9.

Tabel 2. 11 Hasil Prepocessing Data Uji

i	Data uji	Setelah preprocessing	U_i
1	Bagaimana cara mendaftar Autodebet?	bagaimana cara mendaftar autodebet	U_1

Setelah preprocessing masuk ke proses pencarian jawaban dengan melakukan kembali *cosine Similarity* terhadap data uji dengan rule pada sistem tanya jawab.

Berikut adalah hasil penghitungan cosine similarity data uji terhadap rule pada tabel 2.8 dimana data uji disimbolkan dengan U_1 dan aturan disimbolkan dengan R_i.

Tabel 3. 1 Hasil TF.IDF

Term	TF			IDF	W = TF.IDF	
	U_1	R_4	df	Log(n/df)	U_1	R_1
bagaimana	1	1	2	0,544068	0,544068	0,544068
mendaftar	1	1	2	0,544068	0,544068	0,544068
cara	1	1	2	0,544068	0,544068	0,544068
autodebet	1	0	1	0,845098	0,845098	0
di	1	0	1	0,845098	0,845098	0
kuliah	0	1	1	0,845098	0	0,544068

Tabel 3. 2 Hasil Penghitungan

U_1^2	R_1^2	$U_1.R_1$
0,29601	0,29601	0,29601

0,29601	0,29601	0,29601
0,29601	0,29601	0,29601
0,714191	0	0
0,714191	0	0
0	0,845098	0
2,316412	1,733128	0,88803

Tabel 2. 12 Hasil Penghitungan Cosine Data Uji

Data Uji	Aturan	Cosine Similarity
U_1	R_1	0,877971
U_1	R_2	0,329254
U_1	R_3	0,246598

Setelah itu dicari nilai cosine similarity terbesar dari tabel IV.8 maka nilai terbesar itu adalah jawbannya. Berikut jawaban yang dihasilkan oleh sistem.

Tabel 2. 13 Jawaban sistem

No	Pertanyaan (data uji)	Jawaban
1	Bagaimana cara mendaftar Autodebet?	Untuk melakukan autodebet, anda cukup menyetorkan sejumlah uang sesuai dengan jumlah autodebet yang harus dibayarkan ke rekening bank UNIKOM milik anda

3. PENUTUP

Pada bagian ini, akan dibagi menjadi dua bagian yakni kesimpulan dan saran penelitian.

3.1 Kesimpulan

Kesimpulan yang didapat selama pengerjaan peneltian ini mengenai *pendekatan center sentence-based clustering*, didapat bahwa metode *CESED* ini dapat diimplementasikan dalam sistem tanya jawab dengan kasus front office. Dimana berbeda hal dengan metode *clustering* lainnnya, CESED digunakan untuk mengelompokan pertanyaan yang mirip dan membangun rule secara otomatis untuk sistem tanya jawab.

Setelah melakukan pengujian terhadap data uji berikut adalah analisis hasil pengujian terhadap 37 data uji.

1. Terdapat jawaban yang mendekati namun jawaban yang dihasilkan salah seperti pada hasil pengujian No. 32. Pertanyaan yang dimaksud sudah benar maksud jawabannya namun detail dari jawabannya salah.
2. Jawaban yang disediakan didata latih kurang lengkap.
3. Pertanyaan yang mirip bisa menjadi ambigu ketika kata yang terditeksi sistem sama namun pada keterangan nya berbeda seperti “kapan pembayaran kuliah” dengan “kapan pembayaran autodebe” akan menjadi ambigu oleh sistem ketika yang terditeksi hanya “kapan” dan “pembayaran”

Hasil pengujian seperti berikut:

Jumlah jawaban yang benar = 30
Jumlah pertanyaan = 37

$$akurasi = \frac{jumlah\ jawaban\ yang\ benar}{jumlah\ pertanyaan} x100\%$$
$$akurasi = \frac{30}{37} x100\%$$
$$akurasi = 81.08\%$$

Berdasarkan hasil tersebut, sistem tanya jawab diuji dengan menggunnakan 42 data latih didapat akurasi sebesar 81.08%.

3.2 Saran

Bedasarkan hasil analisis dan kesimpulan, terdapat beberapa saran untuk perbaikan pada peneliatian ini, diantaranya adalah sebagai berikut:

1. Pembobotan tidak menggunakan kata tetapi menggunakan w-shingling jadi pembobotan dilakukan berdasarkan input banyak w.

DAFTAR PUSTAKA

[1] Ewerton J. Wantroba and Roseli A.F. Romero, "An Interactive Question-Answer System with Dialogue for a Receptionist Avatar" *SBR Brazilian Robotic Symphony*, 2016.

[2] Romaida Dolarosa Sianturi, "Implementasi Question Answering System dengan Metode Rule-Based pada Banyak Dokumen Berbahasa Indonesia," Institut Pertanian Bogor, 2008.

[3] Thomas Müller and Hinrich Schütze, "Improved Modeling of Out-Of-Vocabulary Words Using Morphological Classes", *Proc. 49th Annu. Meet. Assoc. Computer Linguistic Human Language Technology short Paper*, no. 1, pp. 524–528, 2011.

[4] Shen Song and Yu-N Cheah, "An unsupervised center sentence-based clustering approach for rule-based question answering" *Isc. 2011 - 2011 IEEE Symphony Computer Informatics*, pp. 125–129, 2011.

[5] Zainal A. Hasibuan, "Metodologi Penelitian Pada Bidang Ilmu Komputer Dan Teknologi Informasi", Universitas Indonesia. 2007.

[6] Muhammad Mustaqhfiri, Zainal Abidin, and Ririen Kusumawati, "Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum

Marginal Relevance" *Matics*, vol. 4, no. 4, pp. 134–147, 2011.

[7] Abdul Munif, "*Basis Data*" Direktorat Jendral Peningkatan Mutu Pendidik dan Tenaga Kependidikan, Jakarta. 2013.