Pemberian Peringkat Jawaban pada Forum Tanya-Jawab Online Menggunakan Lexical dan Semantic Similarity Measure Feature
Answer Ranking in Community Question Answering using Lexical and Semantic Similarity Measure Feature

Riska Junia Wulandari[1] , Ade Romadhony[2], Moch. Arif Bijaksana[3]
[123] Prodi S1 Teknik Informatika, Fakultas Informatika ,Universitas Telkom, Bandung
Jalan Telekomunikasi No. 1, Dayeuh Kolot, Bandung 40257
[1] riska.junia@gmail.com, [2] ade.romadhony@gmail.com, [3]arifbijaksana@gmail.com

Abstrak

Maraknya penggunaan internet saat ini membuat banyak forum tanya-jawab (Community Question Answering Online) bermunculan. Bahkan forum tanya jawab yang muncul bukan hanya untuk masalah akademik, tetapi tentang kehidupan disuatu negara seperti QatarLiving Forum. Namun, tidak semua jawaban yang diberikan sesuai dengan pertanyaan yang diajukan. Membutuhkan waktu yang lama untuk menentukan jawaban yang sesuai dengan pertanyaan yang diajukan. Untuk itu, dibutuhkan suatu sistem yang dapat memberikan peringkat pada jawaban untuk membantu merangking jawaban yang sesuai dengan pertanyaan. Adapun tahapan yang dilakukan pada penelitian ini yaitu, dimulai dari preprocessing dataset berdasarkan SemEval 2016 question answering similarity, melakukan ekstraksi fitur untuk membantu proses klasifikasi dengan menggunakan lexical similarity feature, semantic similarity feature, non textual feature grup dan Heuristic. Penelitian ini memfokuskan pada penggunaan lexical similarity dan semantic similarity untuk mencari kemiripan antar pertanyaan dan jawaban. Hasil ekstraksi fitur ini akan dijadikan inputan untuk classfier untuk membuat model yang akan digunakan oleh data uji. Classifier yang digunakan yaitu Support Vector Machine (SVM) dan logistic regression untuk mendapatkan score klasifikasi dimana score ini yang menentukan peringkat sebuah jawaban untuk setiap pertanyaan. Hasil penelitian yang diperoleh menunjukkan pengaruh fitur terhadap kedekatan antara pertanyaan dan jawaban. Lexical similarity feature terutama sub feature Cosine similarity dan LCS menunjukkan semakin tinggi nilai feature pada jawaban semakin jawaban tersebut mendekati pertanyaan yang diajukan. Sedangkan nilai semantic similarity menggunakan Wu Palmer Algorithm, persebaran nilai antar kelasnya lebih merata, sehingga cukup sulit untuk membedakan ciri setiap kelasnya. Non Textual Feature Group membantu dalam melakukan klasifikasi jawaban dan meningkatkan akurasi sebanyak 4%.

Kata Kunci: Community Question Answering, Question Answering System, similarity measure, classifier, peringkat jawaban CQA, MAP.

Abstract

The widespread use of the internet nowadays formed a lot of Online Community Question-Answering. The Online Community Question-Answering not only covers the academic issues, but also about how to living in other country like Qatar Living Forum. However, most of the times the answer that were given did not match the question. It takes some times to choose the right answer that correspond with the answer. For that reason, a rank-based system which can rank the correlation between the question and the answer is needed. The steps that being taken on this study are pre-processing dataset based on SemEval 2016 question-answering similarity, feature extraction to help the classification process using Lexical Similarity Feature, Semantic Similarity Feature, Non-Textual Feature Group, and Heuristic. The classifier used in this study is Logistic Regression and Support Vector Machine (SVM). The value of the classification results are used to rank the most suitable answers to the questions. The results of this study shows the influence of the similarity between the questions and answers. Lexical Similarity Feature, especially Cosine Similarity Sub-Feature and LCS shows that the higher the value of the feature, the closer the answer to the question. On the other hand, the distribution of Semantic Similarity value using Wu Palmer Algorithm in each class is more evenly distributed, so that it is quite difficult to distinguish the characteristics of each class. Non Textual Feature Group is helping to classify the answers and improves 4% of the accuracy.

Keyword: Community Question Answering, Question Answering System, similarity measure, classifier, CQA answer rank, MAP.

1 Pendahuluan

Setiap orang pernah merasakan kebingungan dan ketidaktahuan untuk menjawab pertanyaan yang diajukan kepada mereka, baik pertanyaan dibidang akademik maupun dibidang non akademik. Terkadang dibutuhkan opini, pendapat maupun jawaban dari orang lain untuk membantu menjawab pertanyaan tersebut. Seiring berkembangnya kemajuan teknologi dan banyaknya pengguna internet dimana pengguna internet pada tahun 2015 berdasarkan statistik dari The Statistic Portal bahwa pengguna internet mencapai 3,17 miliar orang[1], mulailah terbentuk banyak forum-forum diskusi online yang dengan nama lain disebut komunitas tanya-jawab (Community Question Answering). Setiap orang dapat bertanya maupun menjawab pada komunitas ini. Sisi positifnya banyak pilihan jawaban yang dapat diambil, namun sisi negatif dari CQA ini terkadang jawaban yang diberikan tidak sesuai atau tidak ada kaitannya dengan pertanyaan, untuk itu dibutuhkan usaha yang lebih untuk mencari jawaban yang sesuai. Banyak kasus dimana jawaban yang diberikan tidak ada kaitannya dengan pertanyaan bahkan berbeda topik pembicaraan, maka dari itu diperlukan suatu sistem yang dapat membantu memberikan peringkat pada jawaban yang paling mendekati pertanyaan yang diajukan.

Pada penelitian ini akan dibuat suatu aplikasi yang dapat memberikan peringkat pada jawaban-jawaban terhadap suatu pertanyaan. Pemberian peringkat berdasarkan tingkat kemiripan antara jawaban dan pertanyaan (Question-Comment Similarity). Untuk melihat tingkat kemiripannya, setiap jawaban dari pertanyaan akan diberikan bobot masing-masing dari hasil ekstraksi fitur. Ekstraksi fitur pada penelitian ini terdiri dari lexical similarity feature, semantic similarity feature dan non textual feature grup dan Heuristic. Namun penelitian ini lebih ditekankan pada similarity measure feature. Pada proses perangkingan jawaban akan digunakan Support Vector Machine (SVM) dan Logistic Regression sebagai classifier dan penentuan peringkat untuk setiap pasangan pertanyaan dan jawabannya. Sistem ini bertujuan untuk melihat pengaruh fitur lexical dan semantic similarity dalam mendeteksi kesamaan antara pertanyaan dan jawaban, selain itu untuk mengetahui fitur-fitur apa saja yang baik dalam pemberian peringkat pada jawaban dan classifier yang menghasilkan nilai evalusi yang terbaik.

2 Dasar Teori dan Perancangan Sistem

2.1 Dataset

Data yang digunakan pada penelitian ini menggunakan dataset SemEval 2016 Task 3 [2] mengenai Qatar Living Forum yang berisi forum tanya jawab tentang kehidupan di Qatar. Dataset ini berbentuk file xml. Terdapat dua dataset yang digunakan yaitu data latih dan data uji. Data yang akan digunakan pada penelitian ini yaitu data pada related question dan pasangan jawabannya (comment). Setiap satu pertanyaan memiliki 10 jawaban yang berada dalam satu thread. Baik pertanyaan maupun jawaban memiliki atribut relevansi antara jawaban dengan pertanyaan yang kelasnya sudah diketahui (apakah jawaban tersebut memiliki kelas good,potentially useful atau bad). Rincian dari masing- masing dataset dapat dilihat pada Tabel 1.

Tabel 1: Statistik Dataset

Kategori	Data Latih	Data Uji
Pertanyaan	1,000	327
Jawaban	10,000	3,270
-good	3,913	1,329
-potential useful	1,620	456
-bad	4,467	1,485

2.2 Ekstraksi Fitur

Ekstraksi fitur (Feature Extraction) digunakan sebagai pengambilan ciri (feature) dari suatu bentuk yang kemudian nilai yang didapatkan akan digunakan dalam proses berikutnya yaitu proses klasifikasi. Tujuan dari penggunaan feature extraction adalah untuk mengambil informasi penting dari data yang diolah.

Pemilihan ekstraksi fitur yang digunakan pada penelitian ini berdasarkan gabungan dari penelitian sebelumnya mengenai pemilihan jawaban pada Community Question Answering SemEval 2015 Task 3, milik pemenang urutan pertama yaitu tim JAIST [1] dan pemenang urutan ketiga yaitu tim QCRI [2]. Namun, ekstraksi fitur yang menjadi fokus utama penelitian ini yaitu Lexical dan Semantic Similarity diambil dari tim QCRI. Terdapat empat fitur utama

[1] http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/. (Diakses tanggal 21 Oktober 2015)
[2] http://alt.qcri.org/semeval2016/task3/index.php?id=data-and-tools

yang digunakan yaitu Lexical Similarity, Semantic Similarity, Non Textual Feature Group dan Heuristic. Adapun penjelasan masing-masing fitur dijelaskan sebagai berikut:

2.2.1 Lexical Similarity

Lexical similarity merupakan pengukuran derajat kemiripan antara dua kalimat yang memiliki kata yang sama. Jika nilai lexical similarity suatu kalimat terhadap kalimat lain bernilai 1, maka kedua kalimat tersebut merupakan kalimat yang sama dan begitu pula kebalikannya, jika nilai lexical similarity antara kedua kalimat bernilai 0, maka tidak ada kata yang sama antara kedua kalimat tersebut [3]. Terdapat tiga sub fitur lexical similarity yang digunakan yaitu sebagai berikut:

1. Cosine Similarity
   Cosine Similarity menghitung nilai kosinus sudut antara dua vektor .Cosine semilariry dapat dihitung dengan persamaan 1.

$$Cosine_sim = \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} (u_i)^2} \times \sqrt{\sum_{i=1}^{n} (v_i)^2}} \quad (1)$$

   Dimana u dan v adalah vektor dari pertanyaan dan jawaban, isinya adalah jumlah kemunculan kata(Term frequency) . n merupakan ukuran dari vektor u dan v. [1].

2. Jaccard Coefficient
   Jaccard coefficient mengukur similarity antara data yang sample set yang terbatas dan didefinisikan sebagai ukuran dari irisan kata yang sama dibagi dengan keseluruhan sample set sebagai berikut:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

   Nilai A adalah kata-kata pada kalimat ke-1 dan nilai B adalah kata-kata pada kalimat ke-2 [4].

3. Longest Common Subsequence

   Longest common subsequence(LCS)dari kelompok A dan B merupakan kelompok terpanjang dari unsur A dan B yang mirip dan berada pada urutan yang sama[3]. Dalam aplikasi biologi sering diperlukan pembandingan DNA dari dua(atau lebih) organisme yang berbeda. Salah satu alasan membandingkan untaian dua DNA ini adalah untuk menentukan kemiripan kedua untaian tersebut sebagai ukuran kedekatan antara kedua organisme tersebut[4].

2.2.2 Semantic Similarity

Perbandingan antara dua kalimat dari pasangan pertanyaan dan jawaban, akan dianggap mirip jika kata kebanyakan kata yang sama atau makna dari pertanyaan dan jawaban itu sama. Semantic similarity antar kata dan sinonim dapat menghasilkan informasi yang berguna ketika pertanyaan dan jawaban memiliki makna/ topik yang sama namun direpresentasikan dalam dalam kata yang berbeda dalam kalimat.

Untuk membandingkan kemiripan dan sinonim suatu kata, maka dibutuhkan suatu kamus atau corpus bahasa sebagai perbandingannya. Perbandingan dapat dilakukan dengan menggunakan WordNet yang kamus kebahasaan dalam bahasa Inggris [2].

2.2.3 WordNet

WordNet merupakan sebuah lexical database yang tersedia secara online dan menyediakan respositori kamus besar lexical dalam bahasa inggris.Wordnet dirancang untuk membangun hubungan antara empat jenis Part of Speech (POS) yaitu noun, verb, kata sifat (adjective), dan kata keterangan. WordNet juga menyediakan algoritma-algoritma untuk menghitung semantic similarity yang disebut WordNet Similarity for Java (WS4J). Algoritma yang digunakan pada penelitian ini yaitu Wu Palmer algorithm [5].

[3] rosettacode.org/wiki/Longest_common_subsequence
[4] https://ecatatan.wordpress.com/2012/12/14/longest-common-subsequence-lcs/comment-page-1/

2.2.4 Wu and Palmer algorithm

Wu palmer menghitung keterkaitan antar dua kata dengan mempertimbangkan kedalam dua synsets pada taksonomi WordNet bersama dengan kedalaman dari LCS (Least Common Subsumer) [5]. Ada pun formula yang digunakan sebagai berikut :

$$\text{Score} = 2 * \text{depth(lcs)} / (\text{depth(s1)} + \text{depth(s2)}) \quad (3)$$

Depth s1 adalah kedalaman dari kata ke 1 dalam wordNet ontology begitu pula depht s2. Score yang dihasilkan dalam rentang nilai 0 sampai 1 ($0 \leq \text{score} \leq 1$).

2.2.5 UMBC Similarity

Terdapat dua pendekatan yang umum digunakan dalam komputasi kesamaan kata yaitu berdasarkan penggunaan thesaurus atau statistik dari corpus yang sangat besar. UMBC semantic similarity melakukan pendekatan hybrid dengan menggabungkan kedua metode tersebut.

Metode statistik yang digunakan, didasarkan pada kesamaan distribusi dan Latent Semantic Analysis(LSA) dan dilengkapi dengan relasi semantic yang di ekstrak dari WordNet. Metode ini mengasumsikan semantic pada frase merupakan komposisi dari komponen kata dan menerapkan algoritma untuk menghitung kesamaan antara dua frase menggunakan word similarity [6].

2.3 Non Textual Feature Group

Non Textual Feature Group merupakan Feature untuk melihat informasi non-textual pada jawaban untuk menentukan kualitas jawaban. Beberapa non textual feature group yang digunakan pada penelitian ini yaitu:

2.3.1 Question Author Feature

Jika penulis pertanyaan dan jawaban berupa orang yang sama, biasanya jawaban tersebut bukan merupakan jawaban yang benar karena kebanyakan jawaban yang dibuat oleh penulis pertanyaan berupa ungkapan terima kasih ataupun pertanyaan lebih lanjut terhadap jawaban-jawaban yang sudah diberikan [1].Fitur ini merupakan boolean fitur yang menghasilkan nilai boolean sebagai keluaran nilainya.

2.3.2 The number of post from the same user

Penelitian ini menyertakan jumlah posting dari pengguna yang sama sebagai fitur karena melalui pengamatan pada data jika satu pengguna memiliki sejumlah jawaban pada thread yang sama, sebagain besar dari jawaban bersifat non-informatif, tidak relevan dengan pertanyaan awal [1]. Sama halnya dengan Question Author Feature, fitur ini merupakan boolean fitur yang menghasilkan nilai 0 dan 1.

2.4 Heuristic

Berdasarkan dataset penelitian, banyak komentar yang baik yang disarankan untuk mengunjungi situs web atau terkandung alamat email. Oleh karena itu, penelitian ini memverifikasi keberadaan url/email dalam jawaban. Karena dengan terdapatnya url/ email dalam jawaban/komentar dianggap berisi informasi yang rinci untuk menjawab pertanyaan [2]. Fitur ini merupakan boolean fitur yang menghasilkan nilai 0 dan 1.

2.5 Evaluasi Performansi Sistem

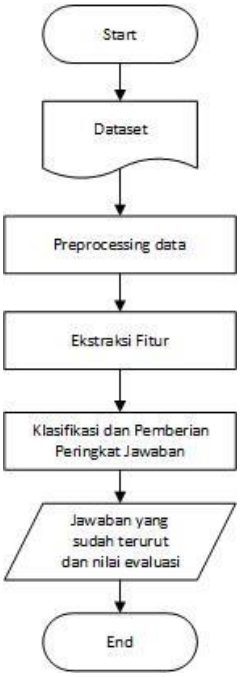
2.5.1 Evaluasi sistem pemeringkatan

Mean Average Precision (MAP) memberikan sebuah nilai tunggal terhadap seluruh titik recall, dari seluruh pengukuran Mean Average Precision dan sudah terbukti dapat menunjukkan tingkat perbedaan dan stabilitas yang baik(Cambridge University Press:2010,pg 159). Mean average precision akan dihitung terhadap sejumlah k dokumen teratas dari dokumen yang di-retrieve dan relevan, dan angkanya akan dirata-ratakan sesuai dengan kebutuhan informasi user. Dokumen yang dianggap relevan pada penelitian ini yaitu jawaban yang memiliki kelas aktual good [7].

$$\text{MAP}(Q) = \frac{1}{|Q|} \Sigma \frac{1}{mj} \Sigma \text{P recision}(R_{jk}) \quad (4)$$

2.6 Perancangan Sistem

Perancangan sistem pembobotan jawaban pada Question Answering, seperti halnya tahapan pada information retrieval, proses ini dimulai dari pengumpulan dataset pertanyaan dan jawaban, preprocessing pada data, ekstraksi fitur dari dataset, dan Perangkingan jawaban. Proses dimulai dari pendefinisian pasangan pertanyaan dan jawaban sebagai masukan, dan keluaran dari proses ini berupa jawaban yang sudah diberikan peringkat berdasarkan kemungkinan jawaban yang paling tepat berdasarkan pertanyaan. Adapun tahapan pembuatan sistem yang dilakukan digambarkan pada Gambar 1.



Gambar 1: Alur Perancangan Sistem.

Terdapat tiga tahapan penting pada penelitian ini yaitu sebagai berikut:

1. Preprocessing Dataset
   Setiap data yang terkumpul pada dataset akan dilakukan preprocessing, untuk data pertanyaan dan juga data jawaban. Preprocessing yang dilakukan yaitu Tokenization, Case Folding, Stopword Removal, dan Stemming. Keluaran yang dihasilkan dari proses ini berupa pertanyaan dan jawaban yang sudah di preprocessing.
2. Ekstraksi Fitur
   Masukan pada proses ini berupa pertanyaan dan jawaban yang sudah dilakukan preprocessing data pada proses sebelumnya. Setiap jawaban akan di beri nilai menggunakan ekstraksi fitur. Jenis-jenis ekstraksi fitur yang digunakan seperti pada Subbab 2.2.
3. Klasifikasi dan Pemberian Peringkat pada Jawaban
   Nilai hasil setiap ekstraksi fitur merupakan parameter inputan yang akan digunakan dalam proses klasifikasi. Selain memberikan class untuk pasangan jawaban dan pertanyaan, classifier akan mencari nilai probabilitas dari setiap jawaban yang akan digunakan sebagai score dalam proses perangkingan jawaban.

   Setelah diberikan peringkat setiap jawaban, peringkat dari setiap jawaban akan dievaluasi menggunakan MAP (Mean Average Precision). Pengujian ini untuk mengetahui apakah jawaban dengan kategori good berada diatas dua kategori lainnya yaitu bad dan potentially useful.

3 Pembahasan

3.1 Implementasi Fitur

3.1.1 Lexical Similarity

Hasil implementasi dari cosine dan jaccard similarity menunjukkan hasil yang sama yaitu semakin tinggi nilai cosine/jaccard similarity pada data latih, semakin jawaban tersebut besar kemungkinan jawaban tersebut berkate-

gori good. Hal tersebut dapat terlihat dari semakin tinggi rentang, semakin banyak pula kelas good dibandingkan kelas bad yang dihasilkan oleh data latih dan data uji.
Sedangkan untuk LCS hasil implementasi menunjukkan semakin panjang untaikan kata yang sama antara jawaban dan pertanyaan, semakin jawaban tersebut mendekati kategori good. Dengan kata lain, semakin banyak karakter huruf yang sama antara jawaban dan pertanyaan, semakin jawaban tersebut memiliki kategori good.

3.1.2 Semantic Similarity

Jika dibandingkan dengan persebaran nilai pada cosine similarity atau jaccard coefficient, nilai semantic similarity lebih merata untuk setiap selangnya baik untuk Wu Palmer Algorithm maupun UMBC Similarity. Jika pada lexical similarity terlihat semakin tinggi nilai lexical dari suatu jawaban, semakin jawaban tersebut dapat dikategorikan Good semakin besar kemungkinannya, namun hasil dari percobaan pada data latih dan data uji dari semantic similarity menampilkan bahwa jumlah persebaran nilai good dan bad hampir setara, sehingga cukup sulit untuk membedakan ciri antara kelas good, bad dan potentially useful.

3.1.3 Non Textual Feature Group

Kedua subfitur dari Non textual Feature Group memberikan pengaruh yang cukup besar terhadap pengklasifikasian jawaban. Pada Question Answer Feature jika penulis pertanyaan merupakan orang yang sama dengan penulis jawaban kebanyakan jawaban tersebut memiliki kelas bad (sekitar 469 jawaban) dan jika penulis jawaban itu bukan penulis pertanyaan itu sendiri, maka kebanyakan jawaban tersebut termasuk dalam kelas good (sekitar 1275 jawaban).

Sedangkan fitur Number of Comment Written by The Same User dapat dilihat bahwa jawaban yang diberikan oleh orang yang sama berpotensi lebih besar pada kelas bad dan jawaban yang diberikan bukan oleh orang yang sama memiliki potensi besar pada kelas good. Hal ini dikarenakan, kebanyakan jawaban yang diberikan oleh orang yang sama dalam satu thread, bukan menjawab pertanyaan diajukan tetapi kebanyakan hanya sebagai spam atau merupakan pertanyaan ulang.

3.1.4 Heuristic

Hasil implementasi menjelaskan bahwa jawaban yang memiliki url banyak terdapat pada kelas good dan potential useful sedangkan data yang tidak terdapat url paling banyak terdapat pada kelas bad. Url berpengaruh terhadap kesesuaian jawaban dengan pertanyaan yang diajukan, karena url memberikan jawaban yang lebih detail dengan memberikan referensi website yang sesuai dengan pertanyaan.

3.2 Hasil Klasifikasi dan Pemeringkatan Jawaban

Berdasarkan hasil analisis terhadap penggunaan logistic regression dan SVM dalam proses klasifikasi dan pemberian peringkat jawaban didapatkan bahwa nilai evalusi yang dihasilkan MAP oleh classifier logistic regression lebih baik dibandingkan nilai yang dihasilkan SVM classfier. Hal ini dapat dilihat pada tabel 2 yang memperlihatkan perbandingan hasil klasifikasi menggunakan logistic regresion dan SVM. Berdasarkan hasil analisis ini,

Tabel 2: Hasil klasifikasi dan Perhitungan Evaluasi

No	Classifier	MAP	F-Measure	Precision	Recall	accuracy
1	Logistic Regresion	71.85%	55.00%	54.20%	59.80%	59.79%
2	SVM	66.73%	53.10%	49.70%	58.00%	57.95%

maka classifier yang akan digunakan untuk menganalisi pengaruh fitur terhadap proses klasifikasi dan pemeringkatan jawaban adalah logistic regression.

3.3 Pengaruh fitur

Penelitian ini menganalisis pengaruh setiap fiturnya melalui kombinasi dari setiap fitur dan nilai yang dihasilkannya. Berdasarkan analisis sebelumnya mengenai analisis hasil klasifikasi, classifier yang menghasilkan MAP terbaik yaitu logistic regression. Adapun hasil pengujian dari kombinasi setiap fitur menggunakan classifier logistic regression seperti pada tabel 3.

Tabel 3: Daftar Kombinasi Fitur

Kombinasi Ke-	Fitur-Fitur yang Digunakan
1	Keseluruhan fitur (Lexical + Semantic Similarity + Non Textual + Heuristic)
2	Lexical Similarity + Semantic + Non Textual
3	Lexical Similarity + Semantic + Url
4	Lexical Similarity + Non Textual + Url
5	Semantic Similarity + Non Textual + Url
6	Lexical Similarity + Semantic Similarity
7	Lexical Similarity + Non Textual
8	Semantic Similarity + Non Textual

Dari kombinasi fitur-fitur pada Tabel 3, Tabel 4 merupakan hasil nilai evaluasi klasifikasi untuk setiap kombinasi fitur menggunakan logistic regression sebagai classifier.

Tabel 4: Hasil Evaluasi Setiap Kombinasi Fitur

Kombinasi Ke-	MAP	F-Measure	Precision	Recall	Accuracy
1	72.19	54.80	51.30	59.70	59.72
2	72.15	55.00	51.40	59.90	59.88
3	64.41	46.40	45.00	52.00	51.96
4	72.11	54.70	51.20	59.60	59.63
5	69.23	55.80	51.90	60.40	60.43
6	64.72	46.20	45.00	51.80	51.83
7	72.05	54.70	51.30	59.70	59.66
8	69.92	56.00	52.10	60.60	60.61

Untuk nilai MAP tertinggi dimiliki oleh kombinasi keseluruhan fitur Lexical Similarity, Semantic similarity ,Non Textual Feature Group dan heuristic dengan nilai 72.19%, Sedangkan untuk nilai evaluasi klasifikasi (Accuracy, F-Measure,Precision dan Recall) tertinggi dimiliki oleh kombinasi fitur Semantic Similarity dan non textual feature group dengan nilai masing-masing 60.61%, 56.00% , 52.10 dan 60.70%.

Untuk mengetahui fitur mana yang memiliki nilai MAP, Accuracy, F-Measure, Precision dan Recall tertinggi, maka dilakukan pengujian terhadap masing-masing fitur dengan menghilangkan fitur tersebut dari kombinasi fitur. Adapun hasil pengujian untuk setiap fiturnya ditampilkan pada Tabel 5.

Tabel 5: Hasil Evaluasi Setiap Fitur

Fitur	MAP	F-Measure	Precision	Recall	Accuracy
Tanpa Cosine Similarity	71.84	55.00	51.40	59.90	59.88
Tanpa Jaccard Coefficient	71.76	54.90	51.40	59.80	59.85
Tanpa LCS	69.11	55.40	51.60	60.00	60.00
Tanpa QA Feature	67.53	50.50	47.80	55.50	55.50
Tanpa Comment ID	70.30	53.70	50.50	58.70	58.69
Tanpa Url	72.15	55.00	51.40	59.90	59.88
Tanpa Wu Palmer	71.95	54.70	51.20	59.60	59.63
Tanpa UMBC	72.27	55.00	51.50	59.90	59.94

Dari hasil pengujian masing-masing fitur pada tabel 5, kombinasi keseluruhan fitur tanpa fitur UMBC similarity menghasilkan nilai MAP tertinggi dengan nilai 72.27%. Jika dilihat baik dari data latih maupun data uji yang digunakan, persebaran nilai pada fitur UMBC semantic similarity,jumlah jawaban yang memiliki kelas bad dan kelas good mendekati seimbang. Hal ini mengakibatkan penggunaan fitur ini tidak terlalu berpengaruh pada nilai evaluasi. Kombinasi akhir yang memiliki nilai MAP tertinggi yaitu Lexical Similarity (Cosine Similarity, Jaccard Coefficient dan LCS), Non Textual Feature Group, heuristic feature dan Wu Palmer Similarity dengan MAP sebesar 72.27%.

4 Kesimpulan

Berdasarkan analisis dan pengujian yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut:

1. Fitur Keterkaitan antar kata (Lexical Similarity) memiliki pengaruh yang lebih tinggi dibandingkan keterkaitan antar makna (Semantic Similarity). Hal ini terjadi karena, persebaran nilai antara kelas good dan bad mendekati seimbang pada fitur Semantic Similarity, Sedangkan pada fitur Lexical similarity terlihat bahwa semakin tinggi nilai lexical suatu jawaban, maka semakin jawaban tersebut mendekati kelas good. Sehingga, fitur lexical similarity memberikan ciri yang lebih jelas terhadap kelas good.

2. Fitur yang memiliki pengaruh tertinggi yaitu Non Textual Feature Group dengan subfitur Question Answer Feature. Fitur ini mendeteksi kesamaan identitass pengguna, antara pembuat jawaban dan pertanyaan (Question Answer Feature). Jika identitas pembuat pertanyaan sama dengan pembuat jawaban, jawaban tersebut kebanyakan memiliki kategori bad. Berdasarkan hal tersebut, fitur ini memberikan ciri yang jelas terhadap kelas bad, sehingga jika terdapat jawaban yang mengandung fitur tersebut kebanyakan merupakan kategori bad. Pengaruh yang diberikan oleh fitur ini sekitar 4%. Sedangkan kombinasi fitur yang menghasilkan nilai MAP tertinggi yaitu Lexical Similarity (Cosine Similarity, Jaccard Coefficient dan LCS), Non Textual Feature Group(Question Answer feature dan number of comment written by the same user), heuristic feature dan Wu Palmer Similarity.

3. Classifier yang menghasilkan nilai MAP tertinggi yaitu Logistic Regression dengan nilai ridge sebesar 40. MAP yang dihasilkan oleh Logistic Regression yaitu 72.27% dengan kombinasi fitur Lexical Similarity (Cosine Similarity, Jaccard Coefficient dan LCS), Non Textual Feature Group(Question Answer feature dan number of comment written by the same user), heuristic feature dan Wu Palmer Similarity.

Daftar Pustaka

[1] Quan Hung Tran, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham. Jaist: Combining multiple features for answer selection in community question answering. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, volume 15, pages 215–219, 2015.

[2] Massimo Nicosia, Simone Filice, Alberto Barrón-Cedeno, Iman Saleh, Hamdy Mubarak, Wei Gao, Preslav Nakov, Giovanni Da San Martino, Alessandro Moschitti, Kareem Darwish, et al. Qcri: Answer selection for community question answeringexperiments for arabic and english. In Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval, volume 15, pages 203–209, 2015.

[3] M Paul Lewis, Gary F Simons, and Charles D Fennig. Ethnologue: Languages of the world, volume 16. SIL international Dallas, TX, 2009.

[4] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Scoring with the jaccard coefficient. Introduction to Information Retrieval, 100:2–4, 2008.

[5] Troy Simpson and Thanh Dao. Wordnet-based semantic similarity measurement. The Code Project. com, 2005.

[6] Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. Umbc ebiquity-core: Semantic textual similarity systems. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics, volume 1, pages 44–52, 2013.

[7] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. Introduction to information retrieval, volume 1. Cambridge university press Cambridge, 2008.