Klasifikasi Jawaban dengan Menggunakan Multiple Features Extraction pada Community Question Answering
Answer Classification using Multiple Features Extraction in Community Question Answering

Bhudi Jati Prio Utomo[1], Moch. Arif Bijaksana [2], Ade Romadhony[3]
[123] Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung
Jalan Telekomunikasi No. 1, Dayeuh Kolot, Bandung 40257
[1] bhudijati@gmail.com, [2]arifbijaksana@gmail.com, [3] ade.romadhony@gmail.com

Abstrak
Berkembangnya Internet semakin memudahkan pengguna dalam pencarian informasi. Community Question Answering (CQA) adalah salah satu sarana yang menyediakan fasilitas tanya jawab dengan mudah dan gratis. Forum diskusi kebanyakan membebaskan pengguna dalam menulis pertanyaan ataupun jawabannya, maka dari itu jawaban-jawaban yang ada pasti sangat bervariasi, terdapat jawaban yang memberikan solusi dan ada juga jawaban yang tidak menjawab pertanyaan. Pada penelitian ini, yang dilakukan berkaitan dengan klasifikasi jawaban terhadap pertanyaan yang ada pada forum diskusi Qatar Living. Identifikasi dilakukan untuk menentukan jawaban mana yang termasuk dalam kelas good, bad, dan potential. Klasifikasi jawaban diselesaikan dengan metode supervised learning. Proses klasifikasi dilakukan pada data yang direpresentasikan oleh fitur seperti Similarity feature (semantic similarity dan cosine similarity), topik model, Textual feature (author), dan Non textual feature (special word, heuristic/link, head word, emoticon, dan question mark). Secara garis besar, terdapat tiga tahap pada penelitian ini yaitu, preprocessing lalu ekstraksi fitur, dan terakhir adalah proses klasifikasi jawaban. Preprocessing terdiri atas tiga tahap yaitu, tokenization, stopword removal, dan lemmatization. Perbedaan penelitian ini dengan penelitian sebelumnya yaitu JAIST adalah dari segi pemilihan fitur. JAIST menggunakan Word matching feature group, Special-component feature group, Non textual feature group, Topic model, Word vector, dan Translation based feature. Berdasarkan hasil evaluasi dari penelitian ini, penulis mendapatkan bahwa klasifikasi yang dilakukan memiliki tingkat akurasi sebesar 72,36 % dan Macro F1 sebesar 54,10 %. Jika dibandingkan dengan hasil SemEval 2015, penelitian ini berada pada urutan ke 3 dari 12 peserta dengan nilai Macro F1 sebagai baseline score untuk pemeringkatannya.
Keyword: community question answering, supervised learning, semantic similiarity, pemodelan topik, qatar living.

Abstract
The growth of the internet makes user to be able to find information easily. Community Question Answering (CQA) is a facility that allow user to do question answering freely and easily. Most discussion forum frees user to write question or answer, therefore the answer is certainly varies, there are answers that provide a solution and does not provide a solution. In this research, we will classify questions that are available in Qatar Living discussion forum. The purpose of identification is to assist whenever we want to specify answer that belong to good, bad, and potential class. Answer classification is performed with supervised learning. Classification process performed on data which represented by feature such as Similarity feature (semantic similarity and cosine similarity), Topic model, Textual feature (author), and Non Textual feature (special word, heuristic/link, head word, emoticon, and question mark). Generally, this research is done with three step: first preprocessing data, then feature extraction, and the last is answer classification process. Preprocessing that is consist of tokenization, stopword removal, and lemmatization. Differences this research between previous research, JAIST is selection feature. JAIST using Word matching feature group, Special-component feature group, Non textual feature group, Topic model, Word vector, and Translation based feature. Based on the evaluation of this research, the author found that classification have been implemented has an accuracy 72,36 % and Macro F1 54,10 %. If compared with the results of SemEval 2015, this research was ranked 3 out of 12 participant with Macro F1 score as a baseline for the ranking.
Keyword: community question answering, supervised learning, semantic similiarity, topic model, qatar living.

1 Pendahuluan

Perkembangan Internet saat ini semakin memudahkan masyarakat dalam pencarian informasi. Sarana di internet tidak hanya menyediakan para pengguna untuk mencari informasi saja, namun sekarang sudah banyak forum diskusi

tanya jawab seperti Qatar Living yang memudahkan pengguna internet untuk berinteraksi dengan media forum sebagai sarana bertukar pikiran dalam mencari suatu solusi. Website Community Question Answering (CQA) seperti Qatar Living telah menyediakan fasilitas yang memungkinkan pengguna untuk memberikan pertanyaan dan jawaban dengan mudah dan gratis. Pada dasarnya fasilitas yang ada di CQA yaitu pengguna membuat thread pertanyaan dan menunggu pengguna lain untuk menjawab pertanyaannya. Pengguna dapat memberikan jawaban dengan bebas namun jawaban yang diberikan sangat bervariasi. Ada jawaban yang memiliki kualitas baik sebagai solusi, buruk, berpotensial dan jawaban yang tidak mengandung informasi apapun. Secara umum hanya beberapa jawaban yang berguna bagi pemberi pertanyaan, ada kemungkinan jawaban-jawaban dari pengguna dapat bersifat spam atau tidak berkaitan dengan pertanyaan dan jawaban yang sama ditulis tidak hanya sekali.

Berdasarkan hal tersebut, penelitian ini bertujuan menganalisis fitur yang cocok dan membuat sistem untuk klasifikasi jawaban ke dalam kelasnya masing-masing. Fitur yang digunakan untuk klasifikasi adalah similarity feature, topik model, textual feature, dan non textual feature. Klasifikasi jawaban dapat diselesaikan dengan menggunakan metode supervised learning. Sistem ini diharapkan dapat memberikan kategori jawaban yang sesuai dengan kelasnya masing-masing, serta dapat menganalisis fitur apa yang cocok dalam klasifikasi jawaban.

2 Dasar Teori dan Perancangan Sistem

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset dari Qatar Living forum yang telah disediakan di Pada SemEval 2015 task 3[1]. Terdapat dua dataset yang digunakan yaitu data latih dan data uji. Penjelasan dari masing masing dataset dapat dilhat pada Tabel 1:

Tabel 1: Statistik English data

Kategory	Training	Testing
Question	2,600	329
Answers	16,541	1,976
Good	8,069	997
Potential	1,659	167
Bad	6,813	812

2.2 Ekstraksi Fitur

Ekstraksi fitur adalah proses ekstraksi dari fitur asli menjadi fitur yang lebih sederhana [1] untuk membantu classifier dalam klasifikasi jawaban pada penelitian ini. Penggunaan fitur dapat memilih term atau kata yang dapat dijadikan sebagai wakil penting untuk kumpulan dokumen. Dalam ekstraksi fitur terdapat beberapa fitur group yang digunakan diantaranya adalah sebagai berikut:

2.2.1 Fitur Similarity

Fitur ini akan mengindentifikasi semua aspek teks yang dapat menunjukan kualitas sebuah jawaban. Fitur ini terdiri atas.

1. Semantic Similarity
   Semantic similarity suatu metode penghitungan untuk mengukur kemiripan kata menggunakan konsep kesamaan semantik. Semantic similarity dapat berguna dalam mengidentifikasikan dua buah objek/kata karena dimana satu set dokumen atau kata saling memiliki kemiripan label semantik dan sinonim dapat menghasilkan informasi yang berguna ketika pertanyaan dan jawaban memiliki makna yang sama namun direpresentasikan dengan kata yang berbeda [2]. Algoritma yang digunakan adalah Wu Palmer dari Word Similarity for Java (WS4J). Berikut

[1] http://alt.qcri.org/semeval2015/task3/index.php?id=data-and-tools

persamaan yang digunakan:

$$WUP(kata_1, kata_2) = -\log \frac{2 \text{ x DepthLCS}}{Depth_1 + Depth_2} \quad (1)$$

DepthLCS merupakan titik temu antara dua kata yang ditinjau pada hirarki WordNet.
Depth1 merupakan kedalaman terpendek pada hirarki WordNet.
Depth2 merupakan kedalaman terpendek pada hirarki WordNet.

2. Cosine Similarity

Cosine similarity merupakan metode similaritas yang paling banyak digunakan untuk menghitung similarity antara dua buah objek/kata berdasarkan leksikalnya [3]. Notasi himpunan yang digunakan untuk menghitung nilai cosine similarity adalah sebagai berikut:

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

keterangan:
u = vektor pertanyaan yang berasal dari jumlah kemunculan kata (term frekuensi) dari pertanyaan.
v = vektor jawaban yang berasal dari jumlah kemunculan kata (term frekuensi) dari jawaban.
u x v merupakan perkalian dot product antara vektor u dengan vektor v
u_i merupakan dimensi dari vektor u dan n adalah ukuran vektor.
v_i merupakan dimensi dari vektor v dan n adalah ukuran vektor.

2.2.2 Fitur Textual

1. Question Mark
Question Mark digunakan untuk mengetahui apakah jawaban yang ada memiliki tanda/simbol tanya. Biasanya jawaban yang terdapat tanda tanya cenderung jawaban yang tidak memberikan suatu solusi, karena penjawab menanyakan hal kembali.

2. Emotikon
Emoticon digunakan untuk mengetahui apakah jawaban yang ada memiliki emotikon seperti simbol tertawa, sedih, dan marah. Biasanya jawaban yang terdapat emotikon cenderung jawaban yang kurang tepat.

3. Special Word
Special word digunakan untuk mengetahui apakah pada jawaban terdapat kata yang spesifik menunjukan kualitas jawaban pada kelas (bad). Contohnya untuk jawaban yang memiliki kata pada kelas bad jawaban biasanya terdapat banyak simbol/kata tawa.

4. Heuristic/Link
Heuristic/Link Feature digunakan untuk mengetahui apakah pada jawaban terdapat alamat website/link. Pada data latih yang ada, biasanya jawaban yang memberikan alamat website/link adalah jawaban yang termasuk dalam kelas good, karena jawaban tersebut memberikan sugesti untuk melihat informasi lebih lanjut.

5. Head Word
Head Word digunakan untuk mengetahui hubungan antara pertanyaan dan jawaban berdasarkan pasangan kata tanya-jawab yang biasanya digunakan. Contohnya, pada pertanyaan terdapat kata tanya where dan jawaban yang baik mengandung kata tempat.

Fitur textual memberikan nilai biner nol dan satu, nol jika jawaban tidak teridentifikasi oleh fitur, dan satu jika jawaban teridentifikasi oleh fitur tersebut.

2.2.3 Fitur Non Textual

1. Question Author Feature
   Question Author Feature digunakan untuk mengetahui identitas pembuat jawaban, apakah jawaban tersebut merupakan milik pembuat pertanyaan atau bukan. Jika penulis pertanyaan dan jawaban berupa orang yang sama, biasanya jawaban tersebut bukan merupakan jawaban yang benar.

2.2.4 Topik Model

Pemodelan topik merupakan pengembangan analisis teks yang bermanfaat dalam pemodelan data tekstual dengan tujuan menemukan topik yang ada pada suatu dokumen. Model yang digunakan adalah model probabilitas Latent Dirichlet Annotation (LDA), dimana berfungsi untuk menemukan korelasi antara kata-kata dengan tema semantik yang tersembunyi di dalam dokumen [4]. Model LDA nantinya mengubah pertanyaan dan jawaban kedalam topik vektor dan menghitung nilai cosine similarity. Pada penelitian ini, pemodelan topik digunakan untuk mengetahui teks atau term apa yang dominan muncul pada suatu pertanyaan atau jawaban.

Latent Dirichlet Annotation (LDA)

Latent Dirichlet Annotation (LDA) adalah merupakan proses generatif sebuah model probabilitas dari data tekstual dimana dapat menjelaskan korelasi antara kata-kata dengan tema semantik yang tersembunyi di dalam suatu dokumen, yaitu topik [5]. Pembahasan tentang pemodelan topik menggunakan LDA telah dilakukan oleh beberapa peneliti sebelumnya, seperti Blei. dkk (2003) [5], Griffiths dan Steyvers (2004) [6] dan, Ponweiser (2012) [7].

Terdapat dua proses dalam topik model, yaitu proses generatif dan proses inferensi menggunakan algoritma gibbs sampling, berikut penjelasannya:

1. Proses Generatif
   Intuisi dibalik LDA adalah suatu dokumen terdiri dari bermacam-macam topik. Secara umum topik terbentuk dari distribusi vocabulary atau kosakata. Sebagai contoh topik genetika memiliki kata tentang genetika dengan nilai probabilitas yang tinggi dan topik evolutionary biology meiliki kata tentang evolutionary biology dengan probabilitas yang tinggi juga. Diasumsikan topik ini telah dispesifikasikan sebelum didapatkan dokumen. Lalu untuk setiap dokumen di dalam koleksi dilakukan ekstraksi kata-kata dengan 2 tahap [5]:
   (a) Secara acak dipilih distribusi atas topik
   (b) Untuk setiap kata dalam dokumen:
      i. Secara acak dipilih sebuah topik dari distribusi atas topik pada langkah 1.
      ii. Secara acak dipilih sebuah kata dari distribusi yang sesuai, atas kosakata.

2. Gibss Sampling
   Pada model generatif LDA kita tidak dapat menemukan variabel tersembunyi. Untuk menemukan variabel tersembunyi ini dapat menggunakan metode inferensi Gibbs Sampling [6]. Berikut formula yang digunakan:

$$P(z_i = j | z_i, w_i, d_i \cdots) = \frac{C^{WT}_{i,j} + \beta}{\sum_{w=1}^{W} C^{WT}_{ij} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \quad (3)$$

   Variabel C^{WT} dan C^{DT} merupakan matriks jumlah dengan dimensi $W \times T$ dan $D \times T$, dimana $C^{WT}_{i,j}$ berisi berapa kali kata w ditetapkan ke topik j dan$C^{DT}_{d_i j}$ berisi berapa kali topik **j** ditetapkan ke sebuah token kata pada dokumen d.

2.3 Evaluasi Performansi Sistem

2.3.1 Evaluasi Klasifikasi

Terdapat beberapa cara untuk mengukur performansi metode klasifikasi diantarnya yaitu dengan menggunakan akurasi, precision, recall dan F-Measure (F1-score). Akurasi adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual, precision adalah tingkat ketepatan antara informasi yang diminta dan jawaban yang diberikan oleh sistem,

sedangkan recall adalah tingkat keberhasilan sistem dalam menemukan kembali informasinya. Jika dilihat pada confusion matrix pada Tabel 2 [8].

Tabel 2: Confusion Matrix

Actual Class	Predicted Class	
	True	False
True	TP	FP
False	FN	TN

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precission} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

Namun terkadang perhitungan antara precision dan recall memiliki perbedaan yang cukup tinggi, untuk itu dilakukan penyetaraan nilai precision dan recall menggunakan F-measure (F1-Score).

$$F1 = 2.\frac{\text{precision.recall}}{\text{precision} + \text{recall}} \tag{7}$$

2.4 Macro F1

Macro-averaged dapat dihitung menggunakan persamaan berikut:

$$L = \{\lambda_j : j = 1...q\} \tag{8}$$

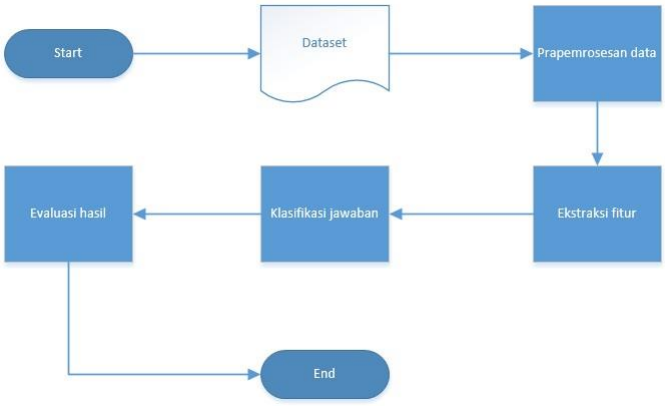
Dimana L adalah semua label, pertimbangan penghitungan evaluasi biner B(tp, tn, fp, fn) yang dihitung berdasarkan jumlah dari true positives (tp), true negatives (tn), false positives (fp) dan false negatives (fn) [9].

$$B_{macro} = \frac{1}{q}\sum_{\lambda=1}^{q} B(tp_\lambda, fp_\lambda, tn_\lambda, fn_\lambda) \tag{9}$$

Pada penelitian ini, penghitungan yang digunakan adalah Macro F1, dimana perhitungan akan menggunakan hasil dari nilai F1 Measure setiap kelas.

2.5 Perancangan Sistem

Secara umum, gambaran sistem yang dibuat dalam penelitian ini terdiri dari proses preprocessing data, ekstraksi fitur, dan klasifikasi. Gambaran umum sistem dapat dilihat pada Gambar 1.



Gambar 1: Gambaran Umum Sistem

Berdasarkan dari Gambar 1, maka alur atau gambaran umum dari sistem ini adalah sebagai berikut:

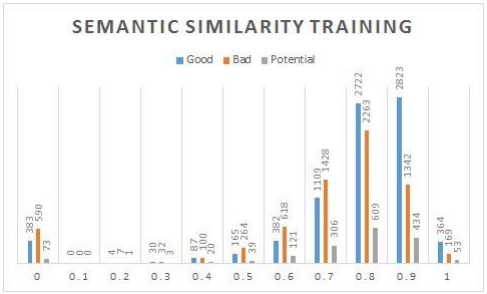
1. Sistem menerima input berupa dataset, dimana dataset ini adalah pertanyaan dan jawaban dari pengguna.
2. Setelah dataset diinputkan, dilakukan proses preprocessing data dengan tahapan tokenization, stopword removal, dan lemmatization.
3. Hasil preprocessing data yang sudah didapat, selanjutnya akan dilakukan tahap ekstraksi fitur.
4. Setelah proses ekstraksi fitur, maka dilakukan klasifikasi untuk menentukan kelas dari jawaban tersebut apakah termasuk di dalam kelas good, bad, dan potential.
5. Setelah proses klasifikasi, maka jawaban yang ada pada setiap kelas akan dievaluasi.

3 Pembahasan

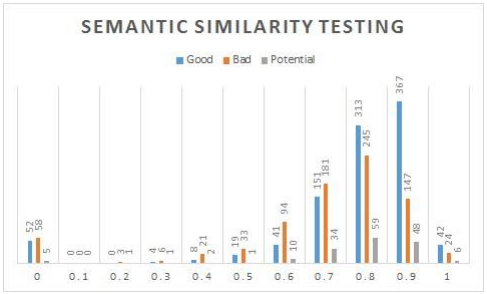
3.1 Implementasi Fitur

3.1.1 Similarity Fitur

1. Semantic Similarity



Gambar 2: Semantic Similarity Data Latih

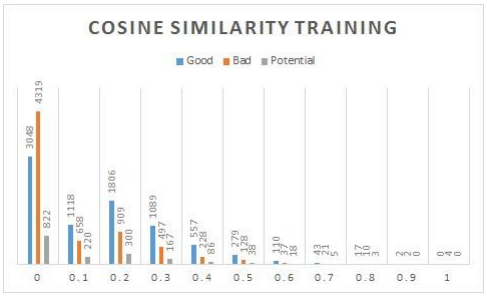


Gambar 3: Semantic Similarity Data Uji

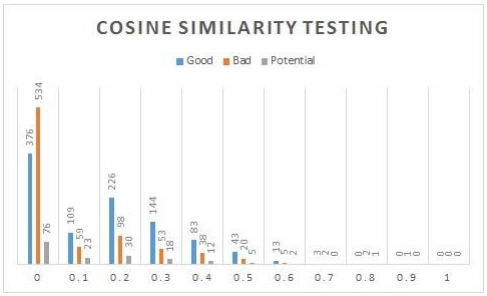
Dari data pada Gambar 2 dan Gambar 3 nilai dari hasil perhitungan semantic similarity tersebar mulai dari 0 hingga 1. Gambar 2 dan Gambar 3 menunjukan persebaran nilai terbanyak pada rentang nilai 0,7 sampai dengan 0,9. Hal ini menunjukan bahwa jawaban yang memiliki nilai semantic similarity yang lebih besar dapat dikategorikan sebagai kelas good. Sedangkan pada nilai 0 kelas bad memiliki frekuensi tertinggi, hal tersebut

menunjukan bahwa jawaban dengan nilai semantic similarity yang kecil dapat dikategorikan sebagai kelas bad. Untuk kelas potential tersebar rata-rata pada rentang nilai 0,6 hingga 0,9, hal tersebut menunjukan bahwa kelas potential berada pada nilai yang sedikit menunjukan bahwa jawabannya terbilang hampir ke dalam kelas good atau berada ditengah-tengah.

2. Cosine Simalarity



Gambar 4: Cosine Similarity Data Latih



Gambar 5: Cosine Similarity Data Uji

Dari data pada Gambar 4 dan Gambar 5 menunjukkan bahwa nilai dari hasil perhitungan cosine similarity tersebar mulai dari 0 hingga 1. Gambar 4 dan Gambar 5 menunjukan persebaran nilai cenderung pada rentang 0 sampai dengan 0,5 dengan kelas bad memiliki jumlah terbanyak pada nilai 0, hal tersebut menunjukan bahwa jawaban yang memiliki nilai cosine yang kecil maka dapat dikategorikan ke dalam kelas bad, sedangkan pada rentang di atas 0 persebaran nilai cosine lebih didominasi oleh kelas good. Hal tersebut menandakan bahwa semakin tinggi nilai cosine maka akan cenderung ke dalam kelas good. Sedangkan untuk kelas potential masih tersebar merata, hal tersebut menjadikan kelas potential menjadi kelas yang sulit untuk diklasifikasi.

3.1.2 Tekstual Fitur

1. Question Mark

Tabel 3: Hasil Fitur Question Mark

Kelas	Data Latih			Data Uji		
	Good	Bad	Potential	Good	Bad	Potential
True	625	2089	253	67	270	26
False	7444	4724	1405	930	542	141

Dari Tabel 3 dapat dilihat bahwa kelas bad memiliki jumlah terbanyak untuk jawaban yang teridentifikasi jawaban tersebut memiliki simbol tanda tanya. Maka dapat disimpulkan bahwa jawaban yang memiliki simbol tanda tanya dapat dikategorikan bahwa jawaban tersebut termasuk kelas bad.

2. Emoticon

Tabel 4: Hasil Fitur Emotikon

Kelas	Data Latih			Data Uji		
	Good	Bad	Potential	Good	Bad	Potential
True	793	1413	223	89	169	19
False	7276	5400	1436	908	643	148

Dari Tabel 4 dapat dilihat bahwa kelas bad memiliki jumlah terbanyak untuk jawaban yang teridentifikasi jawaban tersebut memiliki emotikon. Maka dapat disimpulkan bahwa jawaban yang memiliki emotikon dapat dikategorikan bahwa jawaban tersebut termasuk kelas bad.

3. Special Word

Tabel 5: Hasil Fitur Special Word

Kelas	Data Latih			Data Uji		
	Good	Bad	Potential	Good	Bad	Potential
True	110	468	47	17	46	8
False	7959	6354	1612	980	766	159

Dari Tabel 5 dapat dilihat bahwa kelas bad memiliki jumlah terbanyak untuk jawaban yang teridentifikasi bahwa jawaban tersebut memiliki bad word. Maka dapat disimpulkan bahwa jawaban yang memiliki bad word dapat dikategorikan bahwa jawaban tersebut termasuk kelas bad. Sedangkan untuk jawaban yang tidak memiliki bad word cenderung ke dalam kelas good.

4. Heuristic/Link

Tabel 6: Hasil Fitur Heuristic/Link

Kelas	Data Latih			Data Uji		
	Good	Bad	Potential	Good	Bad	Potential
True	610	446	142	79	113	21
False	7459	6367	1517	918	699	146

Dari Tabel 6 dapat dilihat bahwa kelas good memiliki jumlah terbanyak untuk jawaban yang teridentifikasi bahwa jawaban tersebut memiliki alamat website. Maka dapat disimpulkan bahwa jawaban yang memiliki alamat website dapat dikategorikan bahwa jawaban tersebut termasuk kelas good.

5. Head Word

Tabel 7: Hasil Fitur Head Word

Kelas	Data Latih			Data Uji		
	Good	Bad	Potential	Good	Bad	Potential
True	757	337	96	171	68	22
False	7312	6476	1563	826	744	145

Dari Tabel 7 dapat dilihat bahwa kelas good memiliki jumlah terbanyak untuk jawaban yang teridentifikasi bahwa pertanyaan dan jawaban memiliki pasangan kata tanya-jawab. Maka dapat disimpulkan bahwa jawaban yang memiliki pasangan kata tanya-jawab dapat dikategorikan bahwa jawaban tersebut termasuk kelas good.

3.1.3 Non Textual Fitur

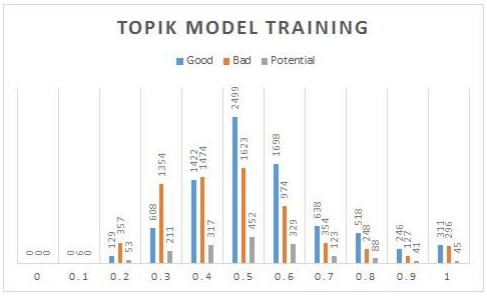
1. Question Author Fitur

Tabel 8: Hasil Fitur Author

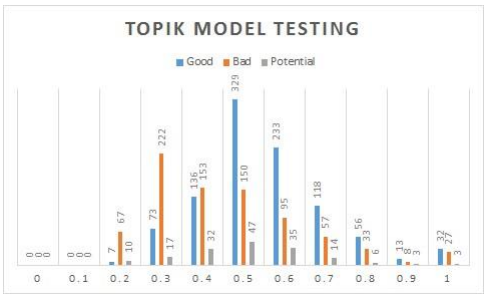
Kelas	Data Latih			Data Uji		
	Good	Bad	Potential	Good	Bad	Potential
True	337	1943	162	32	231	14
False	7732	4870	1497	965	581	153

Dari Tabel 8 dapat dilihat bahwa kelas bad memiliki jumlah terbanyak untuk jawaban yang teridentifikasi bahwa pembuat jawaban tersebut sama dengan pembuat pertanyaan. Maka dapat disimpulkan bahwa jika jawaban dan pertanyaan memiliki user id yang sama dapat dikategorikan bahwa jawaban tersebut termasuk kelas bad.

3.1.4 Pemodelan Topik



Gambar 6: Hasil LDA Data Latih



Gambar 7: Hasil LDA Data Uji

Dari Gambar 6 dan Gambar 7 menunjukkan bahwa nilai dari hasil perhitungan pemodelan topik tersebar mulai dari 0,084 hingga 1. Gambar 6 dan Gambar 7 menunjukan bahwa persebaran nilai pada rentang 0,5 sampai dengan 1 cenderung memberikan kategori jawaban dengan kelas good. Sedangkan pada rentang nilai 0,2 sampai dengan 0,4 cenderung memberikan kategori jawaban dengan kelas bad. Untuk kelas potential tersebar tidak merata yang menyebabkan kelas potential masih sulit untuk diekstraksi. Maka dari itu semakin tinggi nilai cosine similarity dari topik model, maka dapat dikategorikan ke dalam kelas good.

3.2 Hasil Klasifikasi

Analisis hasil klasifikasi dilakukan dengan menganalisis akurasi, precision, recall dan F1-Measure dari hasil klasifikasi tersebut. Proses klasifikasi menggunakan beberapa classifier untuk menguji classifier mana yang menghasilkan nilai akurasi, precision, recall dan F1-measure yang paling tinggi. Analisis pertama menggunakan Support Vector Machine (SVM), kemudian pengklasifikasian kedua menggunakan classifier Bayesian Networks. Berikut analisis hasil perbandingan classifier yang dilakukan:

Tabel 9: Perbandingan Hasil SVM dan Bayesian Networks

Classifier	Macro Precision	Macro Recall	Macro F1
SVM	55,27 %	54,40 %	54,10 %
Bayesian Networks	54,22 %	54,00 %	53,86 %

Berdasarkan hasil pada Tabel 9, maka classifier yang digunakan untuk menganalisis pengaruh fitur terhadap proses klasifikasi adalah SVM, karena pada hasil analisis classifier dalam klasifikasi jawaban, SVM mendapatkan nilai Macro F1 tertinggi.

3.3 Pengaruh fitur

Tabel 10: Daftar Kombinasi Fitur

Kombinasi	Fitur	Macro F1	Akurasi
1	Semantic Similarity	39,21 %	56,68 %
2	Cosine Similarity	40,63 %	58,45 %
3	Heuristic/Link	22,35 %	50,45 %
4	Head Word	22,35 %	50,45 %
5	Special Word	25,98 %	51,92 %
6	Emotikon + Question Mark	42,63 %	63,00 %
7	Semantic + Textual	43,12 %	60,77 %
8	Cosine + Textual	35,60 %	51,41 %
9	Special Word + Emotikon + Question Mark	43,28 %	63,66 %
10	Similarity Feature	40,32 %	57,99 %
11	Topik Model	41,27 %	61,63 %
12	Textual Feature	43,00 %	63,66 %
13	Non Textual Feature	38,00 %	60,52 %
14	Similarity Feature + Topik Model	42,90 %	61,28 %
15	Similarity Feature + Textual Feature	45,95 %	64,72 %
16	Similarity Feature + Non Textual Feature	43,92 %	62,60 %
17	Topik Model + Non Textual Feature	44,62 %	50,70 %
18	Topik Model + Textual Feature	48,01 %	69,48 %
19	Textual Feature + Non Textual Feature	47,80 %	68,72 %
20	Similarity Feature + Topik Model + Textual Feature	49,58 %	66,70 %
21	Similarity Feature + Topik Model + Non Textual Feature	47,62 %	67,25 %
22	Similarity Feature + Textual Feature + Non Textual Feature	51,64 %	70,49 %
23	Topik Model + Textual Feature + Non Textual Feature	49,86 %	71,81 %
24	Cosine + Topik Model + Textual Feature + Non Textual Feature	52,32 %	74,19 %
25	Semantic + Topik Model + Textual Feature + Non Textual Feature	50,61 %	71,60 %
26	Similarity Feature + Topik Model + Textual Feature + Non Textual Feature	54,10 %	72,36 %

Dari data pada Tabel 10 menunjukan hasil dari kombinasi fitur menggunakan SVM classifier. Dapat dilihat bahwa fitur yang paling berpengaruh pada proses klasifikasi jawaban adalah gabungan semua fitur (kombinasi 26). Pengaruh masing-masing fitur dari sebelum penggabungan keseluruhan fitur yaitu:

1. Fitur cosine similarity mempengaruhi nilai Macro F1 sebesar 3,49 % dan lebih besar dibandingkan dengan fitur semantic yang hanya mempengaruhi nilai Macro F1 sebesar 1,80 %.
2. Fitur similarity mempengaruhi nilai Macro F1 sebesar 4,24 % dan nilai akurasi sebesar 0,55 %.
3. Fitur topik model memiliki pengaruh sebesar 2,46 % dari nilai Macro F1 dan mempengaruhi akurasi sebesar 1,87 %.
4. Fitur textual mempengaruhi nilai Macro F1 sebesar 6,48 % dan nilai akurasi sebesar 5,11 %.
5. Fitur non textual memiliki pengaruh nilai Macro F1 sebesar yaitu 4,52 % dan akurasi 5,66 %.

4 Kesimpulan

Berdasarkan analisis dan pengujian yang dilakukan pada Bab 4, maka kesimpulan yang dapat diambil adalah sebagai beikut :

1. Penentuan kemiripan atau keterhubungan kata antara pertanyaan dan jawaban dapat dilakukan dengan fitur similarity. Fitur similarity memberikan pengaruh sebesar 4,24 % dari nilai Macro F1 dan 0,55 % dari nilai akurasi keseluruhan.
2. Hasil klasifikasi jawaban dengan SVM sebagai memberikan nilai Macro F1 yang paling besar.
3. Evaluasi sistem yang terbaik menghasilkan nilai Macro F1 sebesar 54,10 %, dengan hasil tersebut maka penelitian ini telah memenuhi nilai Macro F1 yang ditetapkan sebagai dasar pada SemEval 2015 Task 3 Subtask A English yaitu sebesar 31,23 %.
4. Fitur yang memiliki pengaruh nilai Macro F1 terbesar adalah tekstual fitur diikuti dengan non tekstual fitur, similarity dan, topik model.
5. Fitur semantic similarity masih memiliki pengaruh yang belum terlalu besar. Hal ini disebabkan semantic similarity masih belum mampu untuk mengidentifikasi jawaban yang sesuai dengan pertanyaan yang diajukan, melainkan hanya mengidentifikasi kemiripan pertanyaan dengan jawaban.

Daftar Pustaka

[1] H. Liu and H. Motoda, Feature extraction, construction and selection: A data mining perspective. Springer Science & Business Media, 1998.

[2] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in AAAI, vol. 6, pp. 775–780, 2006.

[3] M. Steinbach, G. Karypis, V. Kumar, et al., "A comparison of document clustering techniques," in KDD workshop on text mining, vol. 400, pp. 525–526, Boston, 2000.

[4] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," IEEE signal processing magazine, vol. 27, no. 6, pp. 55–65, 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[6] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in Proceedings of the 20th conference on Uncertainty in artificial intelligence, pp. 487–494, AUAI Press, 2004.

[7] M. Ponweiser, "Latent dirichlet allocation in r," 2012.

[8] J. H. Martin and D. Jurafsky, "Speech and language processing," International Edition, 2000.

[9] V. Van Asch, "Macro-and micro-averaged evaluation measures [[basic draft]]," 2013.