Klasifikasi Teks untuk Ekstraksi Pasangan Pertanyaan-Jawaban dari *Mega-Thread* di Forum Online

Yudi Wibisono[1)]
Ilmu Komputer, Universitas Pendidikan Indonesia
Jl. Dr. Setiabudhi 229, Lab Basisdata FPMIPA-C, Bandung, Indonesia
email: yudi@upi.edu

Abstrak

Forum online dapat dimanfaatkan dalam pengembangan chatbot dengan mengekstraksi pasangan pertanyaan-jawaban (PJ) dari thread. Pasangan PJ ini akan menjadi sumber pengetahuan untuk chatbot. Makalah ini membahas proses ekstraksi otomatis pasangan PJ dari thread berukuran besar (ribuan posting) yang disebut mega-thread. Klasifikasi teks digunakan untuk menentukan pasangan PJ yang valid dan tidak. Dengan menggunakan 1030 data pasangan quote-tanggapan yang tidak seimbang, akurasi model terbaik diperoleh dengan menggunakan teknik klasifikasi SVM (Support Vector Machine) dengan precision, recall dan F1 kelas minoritas PPJ masing-masing sebesar 0.77, 0,46 dan 0.58. Kinerja model klasifikasi masih memiliki potensi ditingkatkan lebih lanjut dengan penambahan fitur-fitur lain.

Kata kunci: *forum online, mega-thread, klasifikasi teks, pertanyaan-jawaban, SVM*

1. Pendahuluan

Forum online adalah aplikasi web atau mobile yang memungkinkan pengguna membaca, membuat dan mengikuti diskusi atau *thread* secara online. Contoh forum online yang menggunakan bahasa Indonesia adalah Kaskus, DetikForum, dan SerayaMotor. Data pada forum online memiliki potensi untuk dimanfaatkan pada berbagai task pemrosesan bahasa alami karena sifat layanannya yang terbuka, mudah dikumpulkan (di-*crawl*), *posting* dikelompokkan sesuai topik, dan umumnya memiliki moderator sehingga bebas spam.

Forum online umumnya menyediakan beberapa kategori topik. Pengguna dapat memulai diskusi dengan membuat *thread* di tempat yang sesuai dengan kategorinya. Saat ini, muncul pola *thread* baru yang disebut dengan *mega-thread*, yaitu *thread* besar yang dapat berisi ribuan *posting*. *Mega-thread* berisi topik besar atau menjadi wadah diskusi untuk suatu komunitas. Gambar 1 memperlihatkan contoh beberapa *mega-thread* di situs kaskus.co.id pada kategori otomotif dan subkategori roda empat.

Gambar 1. Contoh Mega-Thread di Kaskus

Data posting forum ini dapat dimanfaatkan dalam pemrosesan bahasa alami, terutama untuk chatbot yang saat ini semakin memiliki peranan strategis. Hal ini disebabkan salah satu kegunaan utama forum sebagai media tanya jawab antar penggunanya. Hasil ekstraksi berupa pasangan pertanyaan-jawaban (PJ) dapat menjadi sumber pengetahuan chatbot [1].

Permasalahannya *mega-thread* memiliki karakter berbeda dibandingkan *thread* biasa. *Mega-thread* lebih sulit untuk diproses karena dapat mengandung banyak subtopik dan banyak posting yang sifatnya keluar dari topik. Posting di dalam *mega-thread* dapat berbentuk pertanyaan, jawaban dari pertanyaan, tanggapan dari pertanyaan, pertanyaan balik atau bahkan non pertanyaan seperti opini dan tanggapan. Satu posting juga dapat menjawab sekaligus beberapa pertanyaaan. Ini mempersulit ekstraksi PPJ dari forum. Paper ini membahas penelitian awal ekstraksi otomatis pasangan PJ dari *mega-thread* dengan pendekatan klasifikasi teks.

Untuk Bahasa Inggris, penelitian tentang penggunaan data forum online sudah pernah dilakukan [1][2][3], tetapi belum ada yang membahas mengenai ekstraksi dari *mega-thread*. Sedangkan untuk Bahasa Indonesia, saat ini bahkan belum ada penelitian yang membahas pemanfaatan forum online untuk pemrosesan bahasa alami. Jumlah pengguna internet berbahasa Indonesia yang besar dan semakin berkembangnya *chatbot* membuat penelitian tentang ekstraksi pasangan PJ dari forum menjadi hal yang penting.

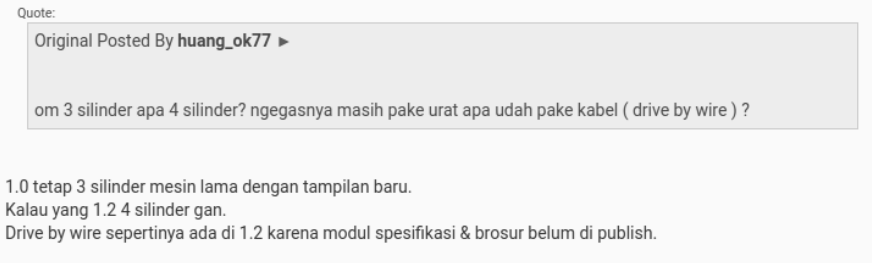
2. Metodologi Penelitian

Dalam bagian ini dibahas mengenai pengumpulan dari forum dan pelabelan data dilanjutkan skenario eksperimen beserta ukuran kinerja model klasifikasi yang digunakan.

2.1. Pengumpulan Data

Data forum dikumpulkan dari forum online Kaskus pada kategori otomotif dan subkategori roda empat. *Mega-thread* yang dipilih adalah thread yang membahas tentang suatu produk mobil yang paling aktif dengan ukuran paling besar. Pengumpulan *posting* dilakukan pada bulan Desember 2017. Jumlah data yang dikumpulkan adalah 32.104 *posting* dalam bentuk HTML.

Di forum Kaskus, saat pengguna membalas suatu *posting* (*reply*) maka posting awal akan muncul sebagai *quote*, dan jawaban pengguna umumnya berada di bawah *quote* (Gambar 2). Oleh karena itu praproses yang pertama dilakukan adalah mengekstraksi *quote* dan tulisan di bawah *quote* dari HTML menjadi teks yang disebut pasangan *quote*-tanggapan. Pasangan ini yang akan menjadi kandidat pasangan pertanyaan-jawaban (PJ).



Gambar 2. Contoh satu pasangan *quote*-tanggapan

Tidak semua pasangan *quote*-tanggapan adalah kandidat pasangan PJ yang valid. Terdapat *quote* yang bukan pertanyaan, dan ada juga tanggapan yang bukan jawaban. Tabel 1 memperlihatkan pasangan *quote*-tanggapan yang bukan merupakan pasangan PJ.

Tabel 1. Pasangan quote-tanggapan yang bukan pasangan PJ

Quote (Q) -Tanggapan (T)	Alasan bukan pasangan PJ
Q:selamat paggiiiii T:Pagi semua...	Quote bukan pertanyaan
Q:itu yang sehat gan.. kalo keluar oli, itu yang namanya bocor.. T:maksudnya uap air kali om, klo iya kyknya tergantung temperatur ruangan deh...	Quote bukan pertanyaan
Q:Foglamp Ayga dan Ayla menggunakan halogen jenis H16 dengan daya 19 watt gan. Coba buka buku manualnya... T:Gan, halogen H16 ada yang warna Kuning ? Kalau ada merek apa ?	Tanggapan berbentuk pertanyaan (bukan jawaban)

Untuk dapat memisahkan pasangan *quote*-tanggapan yang merupakan pasangan PJ dan bukan, maka digunakan teknik klasifikasi teks. Langkah pertama adalah membuat data latih untuk membuat model klasifikasi. Pelabelan dilakukan secara manual oleh satu orang. Dari 1030 pasangan *quote*-tanggapan, diperoleh 189 dengan kelas PPJ (pasangan PJ) dan 841 nonPPJ.

2.2. Skenario Eksperimen

Setelah data latih dikumpulkan pasangan *quote*-tanggapan dibagi menjadi data latih dan data validasi dengan perbandingan 8:2. Teks quote dan teks tanggapan dikonversi menjadi vektor dengan pembobotan TF-IDF. Karena jumlah kelas PPJ dan nonPPJ tidak seimbang (*imbalanced dataset*), maka perlu diperhatikan kinerja model klasifikasi untuk setiap kelas. Ukuran kinerja yang digunakan adalah precision, recall dan F1 untuk setiap kelas.

Beberapa teknik pembelajaran digunakan untuk data ini: Naive Bayes, SVM (Support Vector Machines) dan SGD (Stochastic Gradient Descent). Setelah teknik pembelajaran terbaik didapatkan dari hasil eksperimen, maka eksperimen dilanjutkan dengan mencari parameter yang menghasilkan kinerja terbaik. Tools yang digunakan adalah Scikit-learn [4].

3. Hasil dan Pembahasan

Berdasarkan hasil pelabelan, jumlah instances dengan kelas PPJ ternyata jauh lebih sedikit dibandingkan dengan non PPJ (841 banding 189). Hal ini kontradiktif dengan fakta bahwa forum ini memiliki moderator dan peserta aktif forum umumnya berumur di atas 20 tahun. Rendahnya jumlah PPJ dibandingkan non-PPJ mungkin disebabkan karena forum ini juga digunakan sebagai tempat bersosialiasi dan tempat mengobrol ringan.

Sesuai skenario eksperimen, tiga teknik klasifikasi dicoba untuk dataset ini: Naive Bayes (NB), SVM dan SGD. Tabel 2 memperlihatkan kinerja model Naive Bayes.

Tabel 2. Kinerja Naive Bayes

Kelas	Precision	recall	F1
Non PPJ	0.82	1.00	0.90
PPJ	**0.00**	0.00	0.00

Dapat dilihat pada Tabel 2 bahwa Naive Bayes tidak dapat memprediksi sama sekali kelas pasangan pertanyaan-jawaban (PPJ) sebagai kelas minoritas, walaupun kelas PPJ ini lebih penting untuk diprediksi.

Tabel 3. Kinerja SGD

Kelas	Precision	recall	F1
Non PPJ	0.89	0.97	0.93
PPJ	0.76	0.43	0.55

Tabel 3 memperlihatkan kinerja model SGD. Dapat dilihat bahwa kinerja model jauh lebih baik daripada model Naive Bayes sebelumnya. Precision kelas PPJ menjadi 0.76 dan nonPPJ naik ke 0.89.

Tabel 4. Kinerja SVM

Kelas	Precision	recall	F1
Non PPJ	0.89	0.97	0.93
PPJ	**0.77**	0.46	0.58

Tabel 4 memperlihatkan kinerja teknik SVM. Precision dan F1 untuk kelas PPJ lebih baik daripada SGD. Setelah mencoba beberapa parameter, nilai F1 tidak dapat ditingkatkan lagi.

Tabel 5 memperlihatkan beberapa kesalahan yang dihasilkan oleh model prediksi SVM. Variasi kosa kata yang tinggi karena penggunaan bahasa non formal mungkin menjadi penyebab utama kesalahan prediksi. Untuk penelitian lanjutan, perbaikan kinerja model dapat dilakukan dengan melakukan praproses data seperti normalisasi dan analisis sintaks dan penambahan fitur-fitur lain.

4. Simpulan

Makalah ini membahas penelitian awal untuk mengekstrak secara otomatis pasangan pertanyaan-jawaban dari *mega-thread* di forum online untuk sumber pengetahuan chatbot. Kinerja model cukup memuaskan mengingat model ini masih menggunakan fitur yang sederhana (vektor TF-IDF) tanpa praproses tambahan.

Penelitian selanjutnya perlu lebih mengeksplorasi fitur-fitur lain untuk meningkatkan kinerja model. Ujicoba perlu dilakukan apakah model dapat diterapakan pada domain lain, misalnya satu model klasifikasi dapat diterapkan untuk topik otomotif sekaligus smartphone untuk mengekstrak PPJ (tidak bergantung pada keyword).

Sistem QA atau chatbot juga perlu dikembangkan dengan memanfaatkan PPJ yang sudah diekstrak. Ini diperlukan untuk mempelajari lebih lanjut karakteristik PPJ yang diperlukan oleh sistem QA.

Tabel 5. Kesalahan Prediksi Model SVM

Quote (Q) -Tanggapan (T)	Label	Prediksi
Q: nanti malah TS nya yg pusing gan hahaha oiya mau nanya ayla m sporty bentuk spion nya sama atau beda ama tipe yg lain ya ? T: klo saya perhatikan spion Msporty memang beda ma type X, lebih lebar dikit..mnrt saya lbh cakep jd ga 'kotak' ...	PPJ	nonPPJ
Q: Gan, ane mau ngambil ayla M karena istri ane hamil keguguran melulu klo naek motor pergi pulang kerja. kira2 minusnya apa aja ya gan nih mobil . katanya remnya bunyi2 ya klo di injek dan ada yg berkarat .. ngeri ngeri sedap gan mohon petunjuk T: minus nya 1 gan,,dana buat lahiran (setelah kandungan istri kuat karna naik mobil terus) nanti jadi hilang, ludes,, kidding gan	nonPPJ	PPJ
Q: Gan Gtx gimana hasilnya setelah ganti engine mounting? Getarannya berkurang? T: Sama kyk baru lg gan...Semoga saja sudah di improved sama daihatsunya dan ga getar lagi.	nonPPJ	PPJ

Daftar Pustaka

[1] Cong G, Wang L, Lin CY, Song YI, Sun Y. Finding question-answer pairs from online forums. InProceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval 2008 Jul 20 (pp. 467-474). ACM.

[2] Ding S, Cong G, Lin CY, Zhu X. Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums. InACL 2008 Jun 15 (Vol. 8, pp. 710-718).

[3] Hong L, Davison BD. A classification-based approach to question answering in discussion boards. InProceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval 2009 Jul 19 (pp. 171-178). ACM.

[4] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R. API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238. 2013 Sep 1.