Pengenalan Entitas User Profile Pada Twitter

Entity Recognition of User Profile on Twitter

Titin Pramiyati, Iping Supriana, Ayu Purwarianti
STEI-Institut Teknologi Bandung
Jl. Ganesha 10, Bandung, Indonesia
Email: titin.harsono@gmail.com; iping@informatika.org; ayu@informatika.org

Abstract

Trust scope attribute as an attribute to determine the level of trust resources, will be filled with the data contained in the user profile of Twitter –one of social media-known as Bio Twitter. However, these data should be in accordance with the characteristics and functions of each attribute, such as education attribute must be filled in with the information relating to the educational background of the owner of the profile. To obtain the data corresponding to the trust scope attributes, we perform named entity recognition, which is one of the activities in the process of information extraction. Therefore, this paper describes the results of the entity recognition process performed on data contained in the user profile. Software used to recognize the data as an entity is Indonesia Netagger, which is to perform entity recognition that written in Indonesian language. The software recognizes only five entities namely Person, Organization, Location, Position and Other. We carried out the research by conducting four stages namely entity recognition-with original data-Bio Twitter, error identification, formalizing data, and final test. The results show the success of entity recogniton as follow; Person entity is recognized correctly by 71% of the total data available, the entity Organization recognized correctly by 50%, 20% Position entity recognized correctly, and 50% recognized correctly as Location entity.

Keywords*: trust scope atributes, level trust, user profile, named entity recognition.*

Abstrak

Atribut *trust scope* sebagai atribut untuk menentukan tingkat kepercayaan sumber informasi, akan diisi dengan data yang terdapat pada *user profile Twitter* yang dikenal sebagai *Bio Twitter*. Hanya saja, data tersebut harus sesuai dengan karakteristik dan fungsi dari masing-masing atribut *trust scope*, seperti atribut pendidikan harus diisi dengan informasi yang berkaitan dengan latar belakang pendidikan dari pemilik profil tersebut. Untuk mendapatkan data yang sesuai dengan atribut, kami melakukan *named entity recognition*, yang merupakan salah satu kegiatan pada proses ekstraksi informasi. Oleh karena itu, paper ini menjelaskan hasil proses pengenalan entitas yang dilakukan terhadap data yang terdapat pada *user profile*. Perangkat lunak yang digunakan untuk mengenali data sebagai entitas adalah *IndonesiaNetagger. IndonesiaNettagger*, merupakan perangkat lunak untuk mengenali entitas yang ditulis dalam bahasa Indonesia. Kami melakukan penelitian dalam empat tahap, yaitu pengenalan *entity* dengan data *Bio twitter* yang asli, identifikasi kesalahan proses pengenalan, formalisasi data dan pengujian pengenalan entitas akhir. Hasil penelitian menunjukkan keberhasilan sebagai berikut; entitas *Person* dikenali dengan benar adalah sebesar 71% dari total data entitas yang tersedia, entitas *Organization* dikenali dengan benar sebasar 50%, entitas *Position* 20% dikenali dengan benar, dan 50% entitas *Location* dikenali dengan benar.

Kata kunci: atribut *trust scope*, tingkat kepercayaan, pemilik profil, pengenalan entitas

1. Pendahuluan

User profile adalah tampilan visual dari data pribadi yang dikaitkan dengan pengguna tertentu, dan dapat dianggap sebagai representasi dari sebuah model pengguna. Sebuah *user profile* akan mengacu pada representasi digital identitas seseorang secara eksplisit. Oleh karenanya *user profile* menjadi salah satu layanan yang disediakan pada beberapa layanan internet seperti media sosial.

User profile menyediakan berbagai informasi yang berkaitan dengan jati diri penggunanya, termasuk diantaranya pernyataan tentang tempat bekerja, siapa yang mereka kenal, tempat tinggal, riwayat pendidikan dan sebagainya, umumnya fasilitas ini digunakan untuk berbagi informasi

Received: 26 Februari 2015; Revised: 16 Maret 2015; Accepted: 20 Maret 2015 ; Published online: 30 April 2015

pribadi dengan teman, kerabat, pegawai, atau kepada dunia.

Pengguna internet merupakan sumber informasi yang potensial untuk dimanfaatkan, sehingga ketersediaan informasi tidak hanya berasal dari organisasi resmi saja, akan tetapi masing-masing pengguna dapat mewakili masyarakat untuk berpartisipasi menyediakan informasi yang berkualitas dan dapat dipercaya.

Informasi yang dipercaya dapat diperoleh berdasarkan pada kepercayaan yang dimiliki oleh sumber informasi [1], reputasi sumber informasi [2], dan kepercayaan yang diberikan oleh entiti dengan memperhatikan tingkat kepercayaan (*trust level*) yang dimiliki oleh entiti tersebut [3].

Berbagai model kepercayaan telah banyak dibangun untuk menentukan tingkat kepercayaan, diantaranya model kepercayaan untuk menilai kepercayaan pengguna terhadap aplikasi [4], penentuan kepercayaan pengguna internet [5], dan mekanisme penentuan kepercayaan dan reputasi untuk pengambilan keputusan mendapatkan rekan kerja tanpa harus mengenalnya terlebih dahulu [6].

Terdapat dua jenis penilaian kepercayaan yaitu *direct trust* dan *recommended trust* [7], *direct trust* adalah penilaian kepercayaan berdasarkan pada interaksi langsung yang terjadi, sementara *recommended trust* diperoleh berdasarkan reputasi.

Penilaian kepercayaan langsung, dapat ditentukan berdasarkan *feedback* yang diberikan oleh entiti atau pengguna lain saat terjadi interaksi [1], berdasarkan padakekerapan interaksi yang terjadi [8], dapat ditentukan berdasarkan pada konteks [7] dan *trust scope* [9].

S.Ibotombi Singh dan Smriti K. Sinha [7] menyatakan bahwa kepercayaan dapat dibangun berdasarkan pada *context-sensitive*, *transferable*, *dynamic* dan *history-based*. Penggunaan *context-sensitive* dalam membangun kepercayaan dapat menjadikan suatu *agent* dipercaya pada satu konteks, dan tidak dipercaya pada konteks lain. Sebagai contoh adalah konteks pada layanan pemesanan tiket perjalanan dapat didefinisikan dengan atribut *keberangkatan*, *tujuan*, *tanggal keberangkatan*, dan *kelas*. Sedangkan penggunaan *transferable* dalam menentukan kepercayaan dimaksudkan untuk memberi kepercayaan berdasarkan pada satu konteks untuk memberi kepercayaan pada konteks lainnya, misal seorang yang dipercaya sebagai seorang politisi unggul, akan dipercaya juga sebagai pembicara unggul pula, hal ini karena keterkaitan konteks profesi politisi dengan kemampuannya sebagai pembicara.

Thirunarayan dkk [9] menentukan lingkup kepercayaan atau *trust scope* pada jaringan sosial dan interaksi yang terjadi. Pada kajiannya, Thirunarayan menyatakan bahwa lingkup kepercayaan atau *trust scope* dapat menangkap konteks, kegiatan, fungsi atau domain dari hubungan kepercayaan yang terjadi, misal seorang A mempercayai seorang B karena kemampuan yang dimiliki B, kemampuan B akan menjadi rekomendasi dalam lingkup kepercayaan karena pengetahuan yang dimiliki oleh B. Lingkup kepercayaan yang didasarkan pada pengetahuan yang dimiliki oleh seseorang disebut *referral trust*.

Demikian halnya jika seorang A mempercayai seorang B karena dapat melakukan pekerjaan sesuai dengan lingkup kepercayaannya, maka kepercayaan yang didasarkan pada kemampuan dalam melakukan pekerjaan disebut *functional trust*.

Pada masalah tertentu, seperti penentuan kepercayaan untuk seorang teknisi atau mekanik, *referral trust* dan *functional trust* dapat digunakan secara bersama, karena akan memberikan lingkup kepercayaan yang lebih baik, dibandingkan jika hanya menggunakan salah satu *referral trust* atau *functional trust*. Penggunaan salah satu lingkup kepercayaan akan memberi kemungkinan pemberian kepercayaan yang kurang tepat.

Salah satu properti yang digunakan pada model kepercayaan yang diusulkan oleh Kyounghee Jung adalah properti *interaction significance based on knowledge*. Properti ini digunakan untuk mendapatkan nilai kepercayaan perorangan (*personal trust*), yaitu penilaian yang diberikan oleh pengguna sebuah layanan berdasarkan pada pengetahuan yang dimiliki oleh penggunanya, karena perbedaan pengetahuan akan memberikan nilai yang berbeda untuk tiap pengguna [10].

Selanjutnya, memperhatikan pembahasan penentuan lingkup kepercayaan berdasarkan pada aspek konteks, *transferable*, pengetahuan dan kemampuan dalam menyelesaikan tugas, maka *referral trust* dan *functional trust* dapat mewakili ke-empat aspek tersebut. Penentuan lingkup kepercayaan yang didasarkan pada aspek konteks dan pengetahuan dimasukkan ke dalam *referral trust*, sedangkan *transferable,* kemampuan dan pengetahuan dapat dimasukkan ke dalam *functional trust*.

Untuk menentukan atribut apa saja yang dapat dijadikan sebagai atribut *trust scope*, telah dilakukan survey yang melibatkan 257 responden dan atribut *user profile* dari 4 media sosial yaitu *Facebook*, *Google+*, *Twitter* dan *LinkedIn*. Hasil yang diperoleh dari survey adalah terdapat 8 atribut *trust scope*, yaitu atribut pendidikan, tempat pendidikan, pekerjaan, tempat bekerja, profesi, jabatan, minat dan komunitas.

Twitter adalah layanan jejaring sosial yang lebih dikenal sebagai media sosial yang memungkinkan

penggunanya membuat akun tanpa harus membayar dan dapat mengirim dan membaca pesan teks hingga 140 karakter, yang disebut sebagai kicauan (*tweet*).

Kicauan pengguna *Twitter* dapat terlihat oleh pengguna lain walaupun tidak terdapat ikatan berlanggan dengan cara mengikuti (*follow*) pengguna yang bersangkutan. *Twitter* membedakan pengguna terhadap pengguna yang diikuti sebagai kelompok *following* dan pengguna yang mengikuti adalah kelompok *follower*.

Berdasarkan struktur koneksi pada *Twitter*, berikut kategori utama dari pengguna *Twitter*:

a. Sumber informasi (*Information Source*), pengguna *Twitter* pada kategori ini memiliki pengikut yang besar. Pengguna ini dapat melakukan posting dengan interval umum atau jarang. Meskipun jarang melakukan perubahan informasi, pengguna dengan pengikut yang besar menjadikan perubahan informasi menjadi bernilai. Beberapa sumber informasi juga ditemukan menjadi alat posting berita otomatis dan informasi berguna lain di *Twitter*
b. Teman (*Friends*), pada umumnya hubungan yang terdapat di *Twitter* berada pada kategori ini, dan terdapat berbagai sub-kategori pertemanan yang tersedia, sebagai contoh seorang pengguna dapat memiliki teman, family dan rekan kerja pada daftar pertemanan atau pengikut. Kadangkala pengguna yang tidak kenal dapat menambahkan seseorang sebagai teman
c. Pencari informasi (*Information seeker*), adalah seseorang yang jarang melakukan posting, tetapi mengikuti pengguna lain secara regular.

Sebagai media sosial, *Twitter* juga menyediakan layanan *user profile* yang dikenal sebagai *Bio Twitter*. Berbeda dengan layanan *user profile* pada media sosial lainnya, *Bio Twitter* tidak memisahkan setiap data ke dalam atribut tertentu, sehingga isi dari *Bio Twitter* terlihat seperti sebuah dokumen teks.

Untuk mendapatkan data yang sesuai dengan kriteria atribut *trust scope*, perlu dilakukan pemisahan data yang terdapat pada *Bio Twitter*. Proses pemisahan data ini dapat dilakukan dengan menggunakan proses pengenalan entitas (*Named Entity Recognition*).

Named Entity Recognition (NER) memiliki peran penting pada area aplikasi *Natural Language Processing* (NLP) yang banyak tersedia, seperti ekstraksi informasi, *retrieval information*, tanya-jawab dan peringkasan otomatis. Ciri utama dari tugas NER adalah melakukan identifikasi dan membuat *tag context* pada kata-kata yang tersedia berdasarkan pada kemungkinan kombinasi atas kata-kata tersebut, seperti penentuan panjang minimal kata *word* yang akan diidentifikasi sebagai nama entiti, kata awal, dan sebagainya [11].

Metoda NER diklasifikasikan ke dalam tiga kategori, yaitu *rule-based method*, *statistical-based method*, dan *rule-statistical combined method*. *Rule-based method* menggunakan aturan yang dibuat, dan mengidentifikasi nama entiti yang berbeda dengan cara mencocokan kata dengan aturan yang telah ditentukan. *Statistical-based method* menggunakan korpus yang dianotasi untuk menentukan peluang sebuah kata sebagai nama entiti, jika nilai peluang sebuah kata lebih besar dari nilai *threshold* yang ditentukan, maka kata tersebut akan diidentifikasi sebagai nama entiti. *Rule-statistical combined method*, adalah metoda yang mengkombinasikan antara kedua metoda *rule-based* dan *statistical-based*, seperti penggunaan *rule-conditional random field* (CRF) *combined method* [12].

Named entity seperti *Person*, *Organization*, *Position*, dan *Location* dalam proses identifikasi akan membutuhkan ciri (*feature*) yang merefleksikan properti dari sebuah nama entiti, seperti tipe, kemunculan dan berbagai ukuran umum, baik untuk skala dokumen maupun korpus. Salah satu contoh penggunaan *feature* dalam penentuan nama entiti adalah *feature* kemunculan sebuah kata pada urutan pertama (*first sentence occurence*), karena urutan kemunculan kata dapat menentukan tingkat kepentingan dari kata tersebut [13].

Penelitian yang dilakukan Khodra dan Purwarianti (2013) menggunakan vektor fitur untuk model klasifikasi dan untuk tiap token pada proses ekstraksi informasi dari transaksi *online* di *Twitter*. Vektor fitur yang digunakan pada model klasifikasi menggunakan kategori yang dibangun dengan berdasarkan notasi BIO (*Begin In Other*)-<jenis informasi>, sedangkan vektor fitur untuk tiap token didefinisikan berdasarkan atribut leksikal token tersebut dan tetangganya[14].

Notasi BIO juga digunakan pada sistem NER dengan kerangka kerja CRF (*Conditional Random Field*), untuk memberikan *tag* yang sudah dikenali pada setiap karakter atau kata bahasa China yang terdapat pada dokumen input untuk mengidentifikasi nama entiti [15].

Berdasarkan fungsi *trust scope* dalam menentukan tingat kepercayaan sumber informasi, dan kebutuhan akan kebenaran data yang sesuai dengan karakteristik dari atribut *trust scope*, maka paper ini akan membahas hasil pengujian yang dilakukan pada proses pengenalan entitas untuk merepresentasikan data *user profile* ke dalam atribut *trust scope*, menggunakan *Bio Twitter*

sebagai sumber data dan menggunakan perangkat lunak *IndonesiaNetagger* dalam proses pengenalan entitas untuk sumber data dalam Bahasa Indonesia.

2. Metodologi dan Pengumpulan Data

Pengujian yang dilakukan merupakan proses pengenalan entitas terdiri dari beberapa tahap yaitu pengenalan entiti dengan data *Bio twitter* yang asli, identifikasi kesalahan proses dan pengenalan, formalisasi data dan pengujian pengenalan entitas akhir.

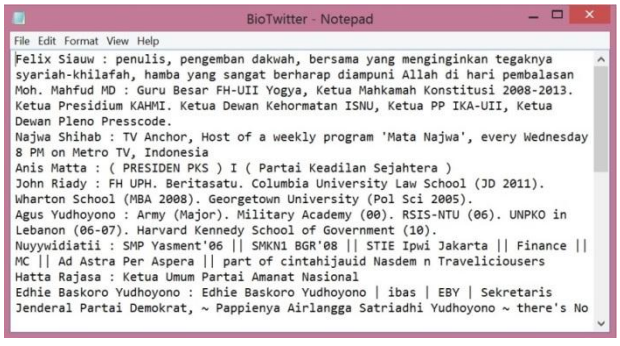
Proses pengenalan entiti terhadap data *Bio Twitter* yang asli dilakukan untuk mengetahui proses pengenalan entitas dapat berjalan dengan baik atau tidak. Hal ini dikarenakan data yang terdapat pada *Bio Twitter* dapat ditulis dengan bentuk yang disukai oleh penggunanya.

Tahap identifikasi kesalahan proses pengenalan dilakukan untuk mengidentifikasi hal-hal yang menimbulkan kegagalan proses pengenalan, dan sekaligus dipakai untuk melakukan perbaikan agar proses pengenalan tidak mengalami kegagalan.

Tahap formalisasi data adalah tahap perbaikan atas sumber data (korpus) yang digunakan berdasarkan hasil identifikasi kesalahan selama proses pengenalan. Tahap terakhir adalah pengujian akhir jika formalisasi data telah selesai dikerjakan, untuk mengetahui hasil pengenalan entitas yang dilakukan.

Proses pengenalan entitas ini menggunakan data *user profile Bio Twitter* yang dikumpulkan melalui pengambilan data *Bio Twitter* secara manual, dan pengambilan dengan memanfaatkan *Aplication Program Interface* (API) yang disediakan oleh *Twitter*.

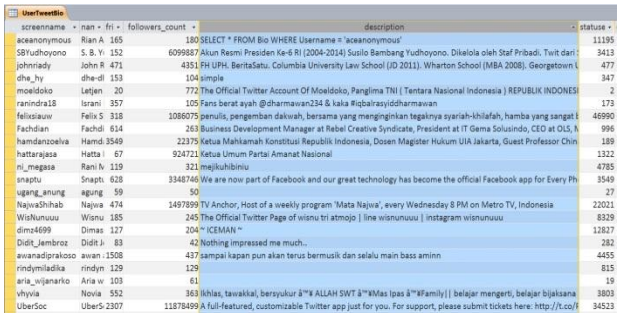
Perangkat yang digunakan pada pengujian proses pengenalan entitas ini adalah perangkat lunak *IndonesiaNETagger* [16] yaitu perangkat untuk melakukan pengenalan entitas untuk korpus Bahasa Indonesia.



Gambar 1. Korpus (BioTwitter.txt).

Pengumpulan data *Bio Twitter* yang dilakukan secara manual berasal dari 20 akun *Twitter* yang dipilih secara acak. Hasil dari pengumpulan data dijadikan sebagai sebuah korpus dengan format file teks (.txt) seperti terlihat pada Gambar 1.

Sedangkan pengumpulan data *Bio Twitter* yang diambil dengan menggunakan API, merupakan *Bio Twitter* dari kelompok *following* dan *follower* dari akun *Twitter* tertentu, data yang berhasil diambil disimpan pada sebuah relasi (*file*) dan diberi nama *UserTweetBio(DBMS:Microsoft Access)*, seperti terlihat pada Gambar 2.



Gambar 2. File UserTweetBio.

Data *Bio Twitter* yang diambil seperti terlihat pada Gambar 3, yaitu akun @anismatta. Data yang tercantum pada *Bio Twitter* terdiri dari teks dan beberapa karakter khusus yang digunakan seperti simbol)|(yang dibuat dari beberapa karakter khusus..

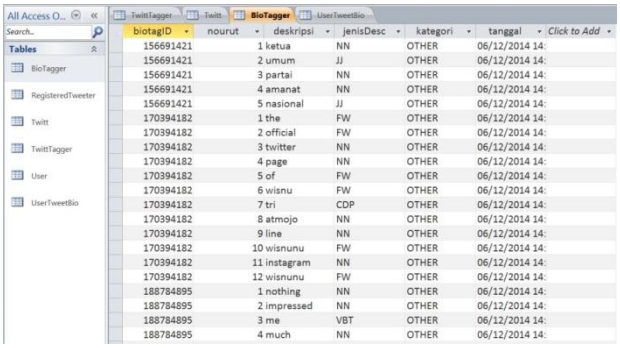
Data *Bio Twitter* yang diambil umumnya menggunakan bahasa Indonesia, akan tetapi ada beberapa data *Bio Twitetr* milik akun tertentu yang menuliskan profesi, nama sekolah dan data lainnya dalam bahasa Inggris.

Gambar 3. *Bio Twitter* pada akun @anismatta.

3. Pengenalan Entitas Menggunakan Perangkat Lunak *IndonesiaNETagger*

Uji coba pertama penggunaan perangkat lunak pengenalan entitas dengan nama *IndonesiaNETagger,* untuk melakukan pengenalan entitas pada data Bio yang terdapat pada relasi *UserTweetBio*, dengan hasil uji coba adalah *IndonesiaNETagger* mengalami kegagalan proses. Langkah penanganan kesalahan proses dilakukan dengan cara menghilangkan semua karakter khusus dan merubah semua huruf pada kata menjadi huruf kecil (*lower case*).

Setelah dilakukan perubahan, ujicoba diulang kembali dengan hasil proses pengenalaan berjalan dengan baik, akan tetapi semua kata dikenali sebagai entitas *Other*, seperti terlihat pada Gambar 4.



biotagID	nourut	deskripsi	jenisDesc	kategori	tanggal
156691421	1	ketua	NN	OTHER	06/12/2014 14:
156691421	2	umum	JJ	OTHER	06/12/2014 14:
156691421	3	partai	NN	OTHER	06/12/2014 14:
156691421	4	amanat	NN	OTHER	06/12/2014 14:
156691421	5	nasional	JJ	OTHER	06/12/2014 14:
170394182	1	the	FW	OTHER	06/12/2014 14:
170394182	2	official	FW	OTHER	06/12/2014 14:
170394182	3	twitter	NN	OTHER	06/12/2014 14:
170394182	4	page	NN	OTHER	06/12/2014 14:
170394182	5	of	FW	OTHER	06/12/2014 14:
170394182	6	wisnu	FW	OTHER	06/12/2014 14:
170394182	7	tri	CDP	OTHER	06/12/2014 14:
170394182	8	atmojo	NN	OTHER	06/12/2014 14:
170394182	9	line	NN	OTHER	06/12/2014 14:
170394182	10	wisnunu	FW	OTHER	06/12/2014 14:
170394182	11	instagram	NN	OTHER	06/12/2014 14:
170394182	12	wisnunu	FW	OTHER	06/12/2014 14:
188784895	1	nothing	NN	OTHER	06/12/2014 14:
188784895	2	impressed	NN	OTHER	06/12/2014 14:
188784895	3	me	VBT	OTHER	06/12/2014 14:
188784895	4	much	NN	OTHER	06/12/2014 14:

Gambar 4. Hasil ujicoba menggunakan sumber data *UserTweetBio.*

Jika semua kata dikenali sebagai entitas *Other*, maka dapat dikatakan bahwa proses pengenalan tidak berhasil, karena proses pengenalan dengan menggunakan *IndonesiaNETagger* harus berhasil mengenali 5 kategori entitas, yaitu *Person*, *Organization*, *Location*, *Position* dan *Other*.

Ujicoba berikutnya menggunakan data korpus *BioTwitter.txt,* yang dibuat secara manual dan belum mengalami perubahan. Ujicoba ini juga mengalami kegagalan proses ketika dilakukan proses pengenalan dengan *IndonesiaNETagger.* Untuk mengetahui sebab kegagalan, dilakukan ujicoba berulang secara manual, selain mengidentifikasi berbagai potensi kegagalan pengenalan, juga dilakukan perubahan pada pada korpus untuk menghindari kegagalan proses.

4. Identifikasi Kesalahan Pengenalan

Berdasarkan pada kegagalan yang terjadi pada proses pengenalan entitas, dilakukan eksperimen dengan menggunakan korpus *(BioTwitter.txt)* yang telah dibuat seperti terlihat pada Gambar 3, ujicoba mengalami kegagalan proses pengenalan. Kegagalan proses bukan disebabkan oleh perangkat lunak *IndonesiaNETagger,* akan tetapi karena file teks yang digunakan berisi data dengan format penulisan seperti tertulis pada *Bio Twitter.*

Format penulisan *Bio Twitter* yang memberikan kebebasan kepada pemilik untuk menuliskan sesuai dengan keinginan mereka, merupakan identifikasi awal yang menyebabkan kegagalan proses, seperti pada contoh berikut ini:

```
@fadlizon Waki Ketua DPR-RI; Wakil Ketua
Umum DPP Partai @Gerindra; Sekjen DPN
@HKTI; @FadliZonLibrary; Dewan Redaksi
HORISON; Ketua @ILUNIFIBUI
```

Untuk mengetahui kesalahan yang menyebabkan kegagalan proses, isi korpus direvisi dan diuji kembali berulang sampai tidak terjadi kegagalan proses. Dari pengujian ini diperoleh beberapa format yang tidak dapat dieksekusi oleh *IndonesiaNETagger*, diantaranya:

a. Kata yang terdapat huruf kapital yang diapit oleh huruf non kapital, seperti PressCode
b. Penggunaan *double hyphen* (--) atau *dash* (–), *slash* (/), *hashtag* (#), dan *pipe* (|)
c. Penggunaan karakter @ yang diikuti dengan kata, seperti @tangandiatas
d. Kata yang mengandung angka setelah huruf, seperti S2, S3
e. Penulisan URL (http://SekolahMonyet.com)

Hal yang sangat menarik pada pengujian ini tidak dikenalinya kata *group*, yang menyebabkan proses NER mengalami kegagalan. Hal ini menjadi unik, karena beberapa kata dalam bahasa Inggris tetap dikenali sebagai entiti walaupun tidak sesuai.

Berdasarkan hasil ujicoba, isi korpus direvisi secara manual, yaitu dengan menghilangkan dan memperbaiki beberapa kata yang menjadi sebab kesalahan, misal pada kata S3 diganti menjadi doktor, kata *group* dirubah menjadi *grup*, dan menghilangkan karakter khusus dengan hasil revisi seperti pada file *BioTwitterCoba.txt*.

Ujicoba kembali dilakukan dengan menggunakan korpus baru yang telah disesuaikan, dan perangkat lunak tidak mengalami kegagalan proses. Berdasarkan hasil dari proses identifikasi kesalahan ini, beberapa bentuk penulisan yang terdapat pada *Bio Twitter* akan dijadikan sebagai dasar dalam proses formalisasi data.

5. Formalisasi Data

Berdasarkan pada hasil identifikasi kesalahan, dilakukan proses formalisasi data yang bertujuan untuk merubah format asli dari *Bio Twitter* menjadi

format yang dapat dieksekusi oleh perangkat *IndonesiaNetagger*. Perubahan ini dilakukan secara otomatis, untuk memenuhi kebutuhan tersebut disiapkan satu program untuk digunakan pada proses formalisasi data.

Perubahan format yang dilakukan pada proses formalisasi data, meliputi proses penghapusan karakter khusus yang terdapat pada *Bio Twitter*. Perubahan yang dilakukan adalah mengganti setiap karakter khusus dengan spasi, dan penggunaan *regular expression* (regex) untuk menangani kesalahan yang disebabkan adanya huruf besar yang diapit oleh huruf kecil pada sebuah kata, seperti terlihat pada penggalan program di bawah ini.

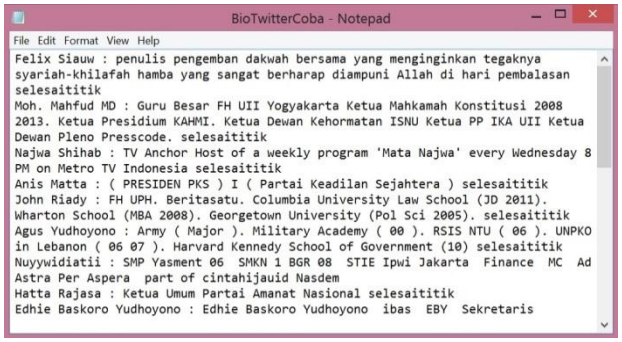
```
String isifile =
FileHelper.openDoc(jTextFieldCorpus.getText());
  isifile = isifile.replaceAll("@", " ");
  isifile = isifile.replaceAll("#", " ");
  isifile = isifile.replaceAll("/", " ");
  isifile = isifile.replaceAll("•", " ");
  isifile = isifile.replaceAll("\\|+", " ");
for (String string : isi) {
if(string.matches("([A-Za-z]+[0-9]+)+_")) {
 String[] s =
string.split("(?<=\\D)(?=\\d)|(?<=\\d)(?=\\D)")
;
string = "";
   for (String splitted : s) {
       string += splitted;
       string += " ";
   }

System.out.println(string+"huruf diikuti
angka");
 }
   if(string.matches("([0-9]+[A-Za-z]+)+"))
System.out.println(string+"huruf diikuti
angka");

if(string.matches("([A-Z]+[a-z]+){2,99}")) {
string = string.charAt(0) +
string.toLowerCase().substring(1,
string.length());

System.out.println(string+"huruf alay");
}
```

Hasil dari perubahan ini dijadikan sebagai korpus baru dalam format teks dengan nama (*BioTwitterCoba.txt*) seperti terlihat pada Gambar 5, sebagai sumber data pada tahap pengujian akhir.



Gambar 5. Hasil formalisasi data

6. Pengujian Akhir

Proses pengenalan enititas menggunakan korpus yang telah diformalisasi berhasil dilakukan, dengan hasil pengenalan entitas *Person* sebesar 71% dikenali dengan benar. Sebuah kata akan dikenali sebagai entitas *Person* jika kata tersebut merupakan sebuah nama. Keberhasilan ini dihitung berdasarkan jumlah kata nama yang dikenali dengan benar dibagi dengan jumlah kata nama.

Sedangkan jumlah kata yang dikenali dengan benar sebagai entitas Organization, sebesar 50% dari 36 kata yang berindikasi nama organisasi, seperti kata Universitas, Partai, DPR dan sebagainya. Seperti pada kalimat Partai Keadilan Sejahtera, pengenalan yang dibenar sebagai entitas *Organization* hanya pada kata Partai dan Keadilan, sedangkan kata Sejahtera dikenali sebagai entitas *Person*.

Hasil uji coba yang terlihat pada Tabel 1, merupakan hasil pengenalan kata yang menunjukkan lokasi sebagai entitas *Location*. Kata yang dikenali dengan benar sebagai entitas Location sebesar 50 %. Umumnya kata yang terdapat pada korpus adalah nama negara dan kota.

Tabel 1. Hasil pengenalan entitas Location

Kata	POS	NER
Yogyakarta	NNP	LOCATION-B
Indonesia	NNP	LOCATION-B
Columbia	NN	OTHER
Lebanon	NNP	LOCATION-B
Jakarta	NNP	LOCATION-B
Bekasi	NNP	PERSON-B
Indonesia	NNP	OTHER
China	NNP	PERSON-I
Beijing	NNP	LOCATION-B
China	NNP	PERSON-B

Tabel 2 menunjukkan hasil pengujian pengenalan entitas untuk kata yang menunjukkan posisi atau jabatan atau entitas *Position*. Keberhasilan pengenalan entitas hanya sebesar 20%, hal ini dikarenakan adanya perbedaan pengenalan antara jabatan dan pekerjaan. Seperti pada kata Dosen atau Professor yang dikenali sebagai entitas *Other* dan *Person*. Karena kata *Dosen* tidak dikenali sebagai jabatan tetapi sebagai pekerjaan atau profesi, sedangkan kata *Profesor* lebih memberi makna sebagai gelar yang melekat pada nama seseorang sehingga dikenali sebagai entitas *Person*.

Tabel 2 Hasil pengenalan entitas Position

Kata	POS	NER
Penulis	NN	OTHER
Guru	NN	OTHER
Ketua	NN	POSITION-B
Ketua	NN	OTHER
PRESIDEN	NN	POSITION-B
Umum	JJ	POSITION-I
Sekretaris	NN	POSITION-B
Jenderal	NN	POSITION-I
Bupati	NNP	OTHER
Konsultan	NN	OTHER
Rektor	NN	PERSON-I
Sekjen	NN	OTHER
Penulis	NN	OTHER
Aktor	NN	OTHER
Sutradara	NN	OTHER
Ketua	NN	PERSON-I
Jurnalis	NN	OTHER
Mahasiswa	NN	PERSON-I
Dosen	NN	OTHER
Professor	NNP	PERSON-I

7. Kesimpulan dan Penelitian Selanjutnya

Data yang terdapat pada *Bio Twitter* dapat digunakan sebagai sumber data untuk mengisi atribut *trust scope*, yang akan digunakan pada penentuan tingkat kepercayaan sumber informasi. Penggunaan *Bio Twitter* sebagai sampel dalam pengujian ini dikarenakan format penulisan pada *Bio Twitter* tidak menggunakan format yang baku dan setiap elemen data tidak dipisahkan sebagai atribut tersendiri seperti pada media sosial *Facebook, Google+ dan LinkedIn*

Sehingga format yang tidak terstruktur ini dapat dijadikan dasar dalam menentukan aturan yang akan digunakan ketika dilakukan proses pengenalan entitas. *IndonesiaNETagger* sebagai perangkat lunak dengan fungsi untuk mengenali entitas untuk data dalam Bahasa Indonesia dapat digunakan dalam proses pengenalan data *Bio Twitter,* walaupun keberhasilan masih belum maksimal untuk 4 kategori entitas.

Berdasarkan hasil ujicoba ini, beberapa penelitian lanjutan sedang dikerjakan, diantaranya adalah pembuatan aturan (*rule*) yang akan digunakan untuk mengenali 8 entitas sesuai dengan atribut *trust scope*. Pembuatan aturan terdiri dari 2 jenis, yaitu jenis pertama pembuatan aturan dengan memanfaatkan hasil pengenalan entitas dengan 4 kategori yang telah tersedia.

Jenis pertama ini akan memanfaatkan entitas *Organization* dalam menentuan entitas Tempat Pendidikan, Tempat Bekerja, dan Komunitas dimana ketiga entitas ini adalah atribut *trust scope*. Sedangkan entitas *Position* dapat dimanfaatkan dalam penentuan entitas Jabatan dan Profesi.

Entitas *Location* dapat juga dimanfaatkan dalam penentuan Tempat Pendidikan, Tempat Bekerja, dan Jabatan.

Sedangkan aturan kedua akan dibuat untuk dapat langsung menentukan entitas sesuai dengan atribut *trust scope*, dan penelitian ini sedang dikerjakan untuk 3 entitas pertama yaitu Pendidikan, Tempat Pendidikan, dan Pekerjaan.

Daftar Pustaka

[1] Y. Gil and V. Ratnakar, "Trusting Information Sources One Citizen at a Time," Proceeding First Int. Semant. Web Conf., 2002.

[2] S. Javanmardi and C. V. Lopes, "Modeling Trust in Collaborative Information Systems," Evolution (N. Y)., 2007.

[3] V. Tundjungsari, J. E. Istiyanto, E. Winarko, and R. Wardoyo, "A Reputation based Trust Model to Seek Judgment in Participatory Group Decision Making," Int. Conf. Distrib. Framew. Multimed. Appl., 2010.

[4] J. Matysiewicz, "Consumer trust – challenge for e-healthcare," Management, pp. 337–342, 2009.

[5] L. Wen, P. Lingdi, L. Kuijun, and C. Xiaoping, "Trust Model of Users ' behavior in Trustworthy Internet ∗ ," Wase Int. Conf. Inf. Eng., pp. 403–406, 2009.

[6] Q. X. Bo Zhang, Yang Xiang, "Trust and Reputation based Model Selection Mechanism for Decision-making," Second Int. Conf. Networks Secur. Wirel. Commun. Trust. Comput., pp. 14–17, 2010.

[7] S. I. Singh and S. K. Sinha, "A New Trust Model based on Social Characteristics and Reputation Mechanisms using Best Local prediction Selection Approach," Int. Conf. New Trends Inf. Serv. Sci., 2009.

[8] E. Marchetti, T. A. Faedo, L. Schilders, and S. Winfield, "Scenario-based testing applied in two real contexts : Healthcare and Employability," 2011.

[9] K. Thirunarayan, P. Anantharam, C. A. Henson, and A. P. Sheth, "Some Trust Issues in Social Networks and Sensor Networks," IEEE, vol. 978–1–4244, pp. 573–580, 2010.

[10] K. Jung and Y. Lee, "Autonomic Trust Extraction for Trustworthy Service Discovery in Urban Computing," 2009 Eighth IEEE Int. Conf. Dependable, Auton. Secur. Comput., vol. 978–0–7695, pp. 502–507, Dec. 2009.

[11] A. Ekbal, S. Saha, and D. Singh, "Ensemble based Active Annotation for Named Entity Recognition," pp. 331–334, 2012.

[12] X. Su, S. Mo, H. Wang, and X. Zhang, "Discovering Significant Persons , Locations and Organizations through Named Entity Ranking," pp. 328–331, 2012.

[13] S. Wang and J. Feng, "A FRAMEWORK FOR ANALYZING THE ' INFORMATION BEARING CAPABILITY ' OF AN INFORMATION SYSTEM," Mach. Learn., no. August, pp. 19–22, 2007.

[14] M. L. Khodra and P. Ayu, "Ekstraksi Informasi Transaksi Online pada Twitter," Cybermatika, vol. 1, no. 1, 2013.
[15] X. Wu, Z. Wu, J. Jia, L. Cai, and C. Science, "ADAPTIVE NAMED ENTITY RECOGNITION BASED ON CONDITIONAL RANDOM FIELDS WITH AUTOMATIC UPDATED DYNAMIC GAZETTEERS Tsinghua National Laboratory for Information Science and Technology (TNList)," pp. 363–367, 2012.
[16] A. F. Wicaksono and P. Ayu, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia."