Identifikasi Hubungan Sebab-Akibat pada Artikel Kesehatan menggunakan Anotasi Elemen Medis dan Paragraf

Susetyo Bagas Bhaskoro[1], Saiful Akbar[2], Suhono Harso Supangkat[3]

***Abstract*—This paper studies natural language processing on medical articles in Indonesian that aims to identify causal relationship and used as public health surveillance information monitoring system. This paper proposes selection-feature conformity, phrase annotation, paragraph annotation, and medical element annotation. System performance evaluation is carried out using intrinsic aprroach which compares supervised classification methods, i.e. naive bayes method and HMM. Results obtained for recall, precission, and f-measure are 0.905, 0.924, 0.910 and 0.706, 0.750, 0.720, respectively.**

***Intisari*—Makalah ini terkait dengan pemrosesan bahasa alami pada artikel medis bahasa Indonesia yang bertujuan untuk identifikasi hubungan sebab-akibat dan digunakan sebagai sistem *monitoring* informasi surveilans kesehatan masyarakat. Usulan pada makalah ini antara lain kesesuaian seleksi fitur, anotasi *phrase*, anotasi paragraf, dan anotasi elemen medis. Evaluasi dari kinerja sistem dilakukan dengan pendekatan intrinsik yaitu membandingkan metode klasifikasi *supervised*, antara lain metode *naive Bayes* dan HMM. Hasil yang diperoleh secara berurutan untuk *recall*, *precision*, dan *f-measure* adalah 0,905; 0,924; 0,910 dan 0,706; 0,750; 0,720.**

***Kata Kunci*— Sebab-akibat, artikel kesehatan, *medical named entities*, seleksi fitur, anotasi elemen medis, anotasi paragraf.**

I. PENDAHULUAN

Pemrosesan bahasa alami menjadi meningkat karena dipengaruhi oleh ketersediaan informasi yang selalu bertambah banyak. Salah satu contohnya adalah pemrosesan bahasa alami pada domain kesehatan yang sumber informasinya tidak hanya berasal dari rekam medis pasien saja, tetapi juga melalui tulisan masyarakat yang dibagikan menggunakan media internet [1], [2]. Tulisan tersebut biasanya membahas tentang pengalaman kesehatan, tanya jawab medis, dan surveilans epidemiologi. Tidak sedikit praktisi, masyarakat, dan unit pelayanan kesehatan yang mencari informasi medis dari internet untuk tujuan tertentu.

[1] *Peneliti, Sekolah Teknik Elektro dan Informatika (STEI) Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40141, INDONESIA (e-mail:susetyo.bagas@gmail.com)*

[2] *Kelompok Keahlian Rekayasa Perangkat Lunak dan Pengetahuan, Sekolah Teknik Elektro dan Informatika (STEI) Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40141, INDONESIA.*

[3] *Kelompok Keahlian Teknologi Informasi, Sekolah Teknik Elektro dan Informatika (STEI) Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40141, INDONESIA.*

Terdapat beberapa penelitian yang memanfaatkan sumber informasi dari internet untuk kebutuhan analisis pada surveilans kesehatan masyarakat dengan memanfaatkan sumber data dari Twitter menggunakan entitas medis, yaitu *disease, symptoms*, *pharmacological*, dan menambahkan *location* untuk mengidentifikasi kemungkinan terjadinya lokasi wabah penyakit [3], [4].

Penelitian lainnya juga memanfaatkan informasi yang terdapat pada Twitter untuk surveilans kesehatan masyarakat dengan entitas medis yang diusulkan, yaitu *sentiment analysis* dan *geocoding* [5], [6]. Referensi [7] menggunakan Twitter untuk mengetahui informasi tentang reaksi obat yang dikonsumsi dengan entitas medis yang digunakan adalah *drug name, brand name, prescribe for*, dan *effect.* Alasan digunakannya *dataset* Twitter adalah untuk melengkapi informasi medis yang tidak didapatkan pada catatan medis secara formal.

Referensi [8] memanfaatkan artikel kesehatan dengan menggunakan entitas medis *drug name*, *ICD Code*, dan *symptom*, begitu juga penelitian yang memanfaatkan artikel kesehatan untuk permasalahan penyakit kanker. Entitas medis yang diusulkan antara lain *substance*, *effect*, *symptom*, *disease*, dan *body part* [9].

Berdasarkan beberapa referensi penelitian sebelumnya, prinsip dasar untuk menghasilkan *labeling* entitas medis disesuaikan dengan *dataset* dan tujuan analisis medis yang akan dihasilkan.

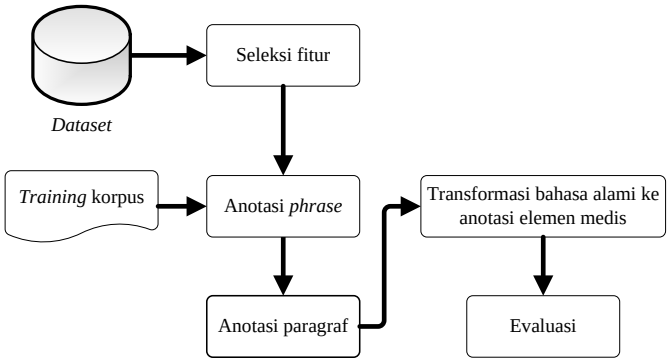
Penelitian yang dilakukan saat ini menggunakan *dataset* artikel medis *online* untuk surveilans kesehatan masyarakat. Tujuannya adalah menampilkan informasi sebab dan akibat, serta proses mitigasi dari kejadian kasus kesehatan tersebut. Informasi sebab-akibat ditampilkan karena pencarian informasi medis akan berakhir apabila telah ditemukan informasi tentang sebab-akibat yang merupakan akar permasalahan pada sebuah kasus kesehatan atau pembahasan inti dari artikel medis [10]. Pengertian dari pola hubungan sebab-akibat adalah kalimat yang dibedakan menjadi dua bagian, yaitu kalimat yang menjelaskan sebab dari sebuah kejadian X dan kalimat lainnya menjelaskan akibat yang menghasilkan kejadian Y yang antara kedua kalimat tersebut saling memiliki keterhubungan. Jika digambarkan pola hubungan sebab $f(x)$ dan akibat atau $f(y)$, maka $f(x,y)$ adalah `<x cause y>` atau `<x → y>` atau `<if x then y>`.

Tantangan pada penelitian ini adalah menentukan elemen medis yang sesuai dengan karakteristik teks pada artikel medis yang kompleks. Usulan kontribusi yang disampaikan adalah (1) memahami pemanfaatan artikel medis *online* untuk

mendukung surveilans kesehatan masyarakat; (2) menentukan seleksi fitur yang paling sesuai untuk digunakan pada karakteristik teks yang kompleks dari artikel medis *online*; (3) menghasilkan anotasi elemen medis untuk digunakan pada bahasa Indonesia; dan (4) menghasilkan anotasi paragraf untuk klasifikasi pembahasan pada sebuah artikel *online*.

II. METODOLOGI

Pada Gbr. 1 diperlihatkan tahapan-tahapan penelitian yang dilakukan.



Gbr. 1 Metodologi penelitian.

A. Dataset

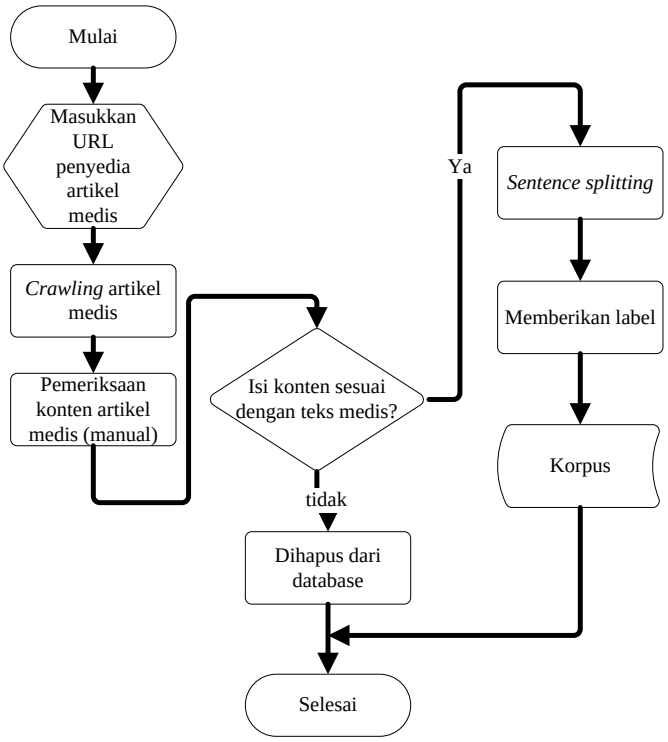
Dataset yang digunakan adalah artikel medis bahasa Indonesia yang diambil dari berita *online*, yaitu kompas.com, detik.com, kumparan.com, viva.com, liputan6.com, metronews.com, dan okezone.com. Konten berita yang digunakan adalah surveilans kesehatan masyarakat.

Karakteristik yang menjadi perhatian dari penggunaan *dataset* dan menjadi analisis awal untuk memastikan bahwa kualitas informasi dari teks medis merupakan fakta dan bukan pemberitaan palsu antara lain sebagai berikut.

- Melakukan analisis judul yang tidak mengandung makna provokatif dan bentuk penulisan kalimat pada konten artikel adalah formal. Aktivitas analisis judul dan konten dilakukan secara manual.
- Mempertimbangkan isi konten artikel dengan melihat hari dan tanggal yang jelas untuk menyatakan waktu kejadian atau peristiwa (*Date, DD-MMM-YYYY*), bukan hanya kata ganti keterangan waktu seperti kemarin, dua hari yang lalu, dan sebagainya.
- Isi konten pada artikel memiliki kutipan dari narasumber, dan bukan hanya opini pribadi dari penulis saja.
- Sumber berita yang digunakan berasal dari situs berita yang telah terverifikasi.

Gbr. 2 adalah diagram alir tahapan pengumpulan artikel medis dari media *online*. Jumlah artikel medis yang disimpan di basis data sebanyak 500 buah. Selanjutnya, dari masing-masing artikel tersebut dilakukan pemrosesan teks *sentence splitting* dan didapatkan kalimat sebanyak 10.176 buah dan kata sebanyak 212.264 buah.

Pada tahapan selanjutnya, masing-masing kalimat diberi label, antara lain anotasi frasa, anotasi paragraf, dan anotasi elemen medis. Data yang disiapkan untuk tahapan data latih dan data uji memiliki komposisi awal sebesar 40% atau sebanyak 200 kalimat untuk data latih dan 60% atau sebanyak 300 kalimat untuk data uji. Komposisi yang dimaksud adalah suatu kondisi ketika sistem belum melakukan pengujian pemrosesan teks. Ketika sistem telah melakukan pemrosesan teks, maka jumlah komposisi data latih akan selalu bertambah menyesuaikan *reward* dan *punishment* yang diberikan oleh seorang pakar ke *dataset*. Proses *reward* berfungsi ketika seorang pakar memberikan masukan positif ke dalam sistem, dan selanjutnya sistem melakukan pembaruan data ke dalam korpus. Ketika proses *punishment* diberikan oleh seorang pakar, maka sistem tidak akan melakukan penambahan data.



Gbr. 2 Diagram alir pada tahapan pengumpulan artikel medis dari media *online*.

B. Seleksi Fitur

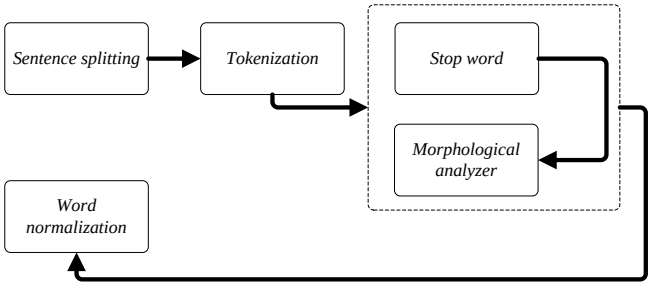
Pemilihan fitur disesuaikan dengan karakteristik teks pada artikel medis yang kompleks dan makna kalimat medis. Tabel I menyajikan seleksi fitur yang digunakan pada penelitian ini.

TABEL I
SELEKSI FITUR

Fitur	Penjelasan
Prapemrosesan	Perbandingan antara menggunakan dan tanpa menggunakan prapemrosesan. Hasil yang lebih baik adalah tanpa menggunakan prapemrosesan.
N-gram	Jumlah *n-gram* yang digunakan pada penelitian ini adalah *tri-gram.*
Anotasi frasa	Melakukan identifikasi makna positif dan makna negatif pada sebuah kalimat.

Hasil analisis penggunaan fitur prapemrosesan pada artikel medis adalah keluaran lebih optimal ketika tidak menggunakan fitur tersebut dibandingkan dengan mengguna-

kan fitur prapemrosesan, karena hasilnya menjadi ambigu. Gbr. 3 menunjukkan tahapan prapemrosesan yang dimaksud.



Gbr. 3 Tahapan prapemrosesan.

Tahap pertama adalah *sentence splitting*, yaitu melakukan pemisahan kalimat sampai dengan tanda *delimiter* '.'. Namun, tidak semua tanda *delimeter* adalah penanda akhir dari sebuah kalimat. Contohnya seperti 'dr.'. Kata tersebut bukan akhir dari sebuah kalimat, melainkan gelar profesi.

Tahapan kedua adalah tokenisasi, yaitu memisahkan setiap kata dari sebuah kalimat. Contoh permasalahannya seperti frasa "rumah sakit". Frasa tersebut adalah satu kesatuan, bukan "rumah" dan "sakit". Begitu juga dengan "Kementerian kesehatan". Frasa tersebut adalah satu kesatuan, bukan "kementerian" dan "kesehatan".

Tahapan yang terakhir adalah *morphological analyzer*, yaitu mengubah sebuah kata menjadi kata dasar. Contohnya adalah "diimunisasi". Kata tersebut memiliki imbuhan (di-), (sasi-), dan kata dasarnya adalah "imun". Diimunisasi memiliki makna diberikan vaksin untuk kekebalan tubuh, sedangkan imun adalah sistem kekebalan tubuh.

Analisis pada seleksi fitur lainnya adalah *n-gram*, seperti yang terlihat pada (1).

$$P\ (k_1, k_2, \dots, k_n\ ) = P(k_1)P(k_2|k_1)\dots P(k_n \mid k_1 \dots\ k_{n-1}). \quad (1)$$

Penggunaan *n-gram* pada di sini menggunakan n = 3, sehingga disebut dengan trigram, dan persamaannya seperti pada (2).

$$P\ (k_n|k_1, \dots, k_{n-1}\ ) = P(k_n \mid k_{n-2} \dots\ k_{n-1}). \quad (2)$$

Persamaan sebelumnya dapat dituliskan seperti berikut.

$$P\ (k_1, k_2, \dots, k_n\ ) = P(k_1)P(k_2|k_1)\prod_{i=3}^{n} P(k_i \mid k_{i-2} \dots\ k_{i-1}). \quad (3)$$

Tahapan selanjutnya adalah mengadopsi penggunaan *semantic analysis* seperti pada [5] menjadi anotasi frasa untuk mengidentifikasi makna positif atau negatif. Analisis opini menggunakan analisis kalimat positif (+) untuk mewakili makna sebab, dan analisis kalimat negatif (-) untuk mewakili makna akibat, sehingga hasil sintaksis yang didapatkan adalah seperti berikut.

```
<Demam Berdarah Dengeu (DBD)> / NN
kembali  / SC
<memakan korban jiwa> / VB + (NEG)
di / IN
Kota Pekanbaru. / NNP
Seorang bocah perempuan / NNP
berusia / CD
tujuh tahun / CD
bernama / DT
Nursabrina / NNP
<meninggal dunia> / VB + (NEG)
akibat / SC
<gigitan nyamuk Aides Aigepty>. / VB + (NEG)
```

Hasil eksplorasi kalimat tersebut menampilkan kalimat "memakan korban jiwa" / VB + (NEG) dan "meninggal dunia" / VB + (NEG) memiliki makna negatif.

C. Anotasi Paragraf

Karakteristik teks pada artikel *online* adalah kompleks atau *multiple sentences*, yaitu sebuah artikel panjang yang biasanya memiliki beberapa paragraf dan penjelasannya dimulai dengan latar belakang, kemudian menyampaikan masalah dan memberikan solusi pada bagian inti artikel, serta diakhiri dengan kesimpulan atau pesan yang dianggap penting.

Posisi dari pola hubungan sebab-akibat memungkinkan berada di paragraf yang berbeda. Artinya, kalimat sebab dapat saja dituliskan di awal paragraf dan kalimat akibat dijelaskan di akhir paragraf, begitu juga sebaliknya.

Tugas utama dari anotasi paragraf adalah menghasilkan intisari kategori kata pada konten medis yang terdiri atas paragraf inti, paragraf pendukung, dan paragraf kesimpulan. Daftar anotasi posisi paragraf seperti yang diperlihatkan pada Tabel II.

TABEL II
ANOTASI POSISI PARAGRAF

Anotasi paragraf	**Karakteristik**

Paragraf inti	Karakteristik yang dimiliki antara lain:
	1. Kalimat yang memiliki frekuensi kemunculannya lebih sering dibandingkan dengan kalimat lainnya dan menggambarkan kejadian yang sedang dibahas.
	2. Kalimat inti mengandung makna akibat.
	3. Kalimat yang menjelaskan tentang fakta dari suatu kejadian yang dibahas.
	
	Elemen yang berhubungan dengan paragraf inti antara lain:
	1. Jumlah dalam angka: jumlah kejadian, bukti kejadian.
	2. Akibat.
	3. Farmasi.
	4. Diagnosis.
	5. Tanda di dalam tubuh terkait dengan gejala.
	
	Teknik penggunaannya adalah melakukan perhitungan frekuensi jumlah kemunculan kata dan melakukan *ranking* berdasarkan nilai *threshold* pada sebuah artikel.
	`Loop artikel, jumlah[]`
	`  {element_paragraf_inti,`
	`  posisi_paragraph_inti = n`
	`  n = n+1}`
	`End loop`

Paragraf penjelas	Karakteristik yang dimiliki antara lain: 1. Kalimat yang identik dengan informasi penyebab dari sebuah kejadian.

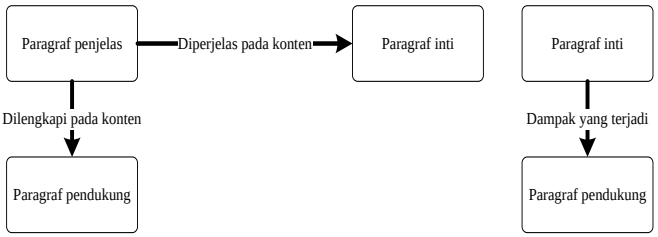
TABEL II (LANJ.)
ANOTASI POSISI PARAGRAF

Anotasi paragraf	Karakteristik

Paragraf penjelas	2. Kalimat yang menjelaskan tentang subjek yang dibahas pada artikel tersebut.
	3. Kalimat yang mendefinisikan nama penyakit yang dibahas pada artikel tersebut;.
	4. Kalimat yang menggambarkan keterangan tempat.
	
	Elemen yang berhubungan dengan paragraf pendukung antara lain:
	1. Lokasi: negara, kota, kabupaten, alamat.
	2. *Disease* (penyakit).
	3. Sebab.
	4. Subjek: orang, benda.
	5. Kalimat perbandingan.
	
	Teknik penggunaannya adalah melakukan perhitungan frekuensi jumlah kemunculan kata dan melakukan *ranking* berdasarkan nilai *threshold* pada sebuah artikel.
	`Loop artikel, jumlah[]`
	`{element_paragraf_inti,`
	`posisi_paragraph_inti = n`
	`n = n+1}`
	`End loop`

Paragraf kesimpulan	Karakteristik yang dimiliki antara lain:
	1. Kalimat yang menjelaskan tentang mitigasi dari kejadian yang sedang berlangsung.
	2. Kalimat yang mendefinisikan tentang fasilitas yang dibahas pada artikel tersebut.
	
	Elemen yang berhubungan dengan paragraf kesimpulan antara lain:
	1. Mitigasi.
	2. Fasilitas yang diberikan.
	
	Teknik penggunaannya adalah melakukan perhitungan frekuensi jumlah kemunculan kata dan melakukan *ranking* berdasarkan nilai *threshold* pada sebuah artikel.
	`Loop artikel, jumlah[]`
	`{element_paragraf_inti,`
	`posisi_paragraph_inti = n`
	`n = n+1}`
	`End loop`

Jika pola hubungan sebab-akibat digambarkan antara hubungan anotasi paragraf, yaitu paragraf inti, paragraf penjelas dan paragraf kesimpulan, adalah seperti ditunjukkan pada Gbr. 4.



Gbr. 4 Anotasi paragraf.

Gbr. 4 menunjukkan bahwa pola hubungan sebab-akibat dapat diidentifikasi apabila salah satu kelas sebab atau akibat hasil klasifikasi dari anotasi elemen medis terdapat pada hubungan anotasi `paragraf penjelas → paragraf inti`, `paragraf penjelas → paragraf pendukung`, dan `paragraf inti → paragraf pendukung`.

D. Transformasi Bahasa Alami ke Anotasi Elemen Medis

Usulan pada makalah ini adalah melakukan transformasi setiap bahasa alami ke anotasi elemen medis yang sudah ditentukan. Hasil anotasi tersebut digunakan untuk melakukan analisis identifikasi sebab-akibat. Anotasi elemen medis yang dimaksud seperti yang disajikan pada Tabel III.

TABEL III
ANOTASI ELEMEN MEDIS

Anotasi Elemen medis	Penjelasan
Lokasi	Digunakan untuk melakukan ekstraksi lokasi dan nama tempat tentang kejadian atau peristiwa yang sedang dibahas. `Lokasi (x,y) → probabilitas lokasi.` $P(c \mid x) = \frac{P(x \mid c) \times P(c)}{P(x)}$
Populasi penyakit	Digunakan untuk melakukan ekstraksi isu dan masalah penyakit. `Populasi_penyakit (x,y) → probabilitas biological named entity recognition (biological NER).` $P(c \mid x) = \frac{P(x \mid c) \times P(c)}{P(x)}$
Akibat	Digunakan untuk melakukan ekstraksi dampak yang dirasakan oleh subjek karena populasi penyakit tertentu. `Akibat (x,y) → Populasi_penyakit(x), Lokasi (y)`
Jumlah kejadian	Digunakan untuk melakukan ekstraksi jumlah statistik dalam angka yang berasal dari populasi penyakit tertentu. `Jumlah kejadian (x,y) → Akibat (x,y)`

Sebab	Digunakan untuk melakukan ekstraksi penyebab utama. Metode yang digunakan untuk mendapatkan sebab adalah menghitung statistik kemunculan kata pada artikel *online* dan klasifikasi populasi penyakit tertentu.
	`Sebab_1 (x,y) → Populasi_penyakit(x,y)`
	`Sebab_2 (x,y) → Jumlah_kejadian (x,y)`
	`Sebab_3 (x,y) → jumlah_kejadian(x), lokasi(y)`

Fasilitas	Digunakan untuk melakukan ekstraksi nama institusi yang memberikan penanganan berdasarkan akibat yang ditimbulkan dari populasi penyakit tertentu.
	`Fasilitas_1 (x,y) → Populasi_penyakit(x), jumlah_kejadian(y)`
	`Fasilitas_2 (x,y) → Akibat(x,y)`
	`Fasiltias_3 (x,y) → Sebab_1(x,y)`

TABEL III (LANJ.)
ANOTASI ELEMEN MEDIS

Anotasi Elemen medis	Penjelasan

Pemulihan	Digunakan untuk melakukan ekstraksi solusi yang disediakan dan dilakukan dari fasilitas di lokasi tertentu.
	
	`Pemulihan_1 (x,y) → Fasilitas_1 (x,y)`
	`Pemulihan_2 (x,y) → Fasilitas(x), lokasi(y)`

Penanggung jawab	Digunakan untuk melakukan ekstraksi subjek yang bertanggung jawab untuk permasalahan populasi penyakit tertentu dan pemulihan yang dilakukan oleh fasilitas di lokasi tertentu.
	
	`Penanggung_jawab_1 (x,y) → Fasilitas_1 (x,y)`
	`Penanggung_jawab_2 (x,y) → Sebab_3 (x,y)`
	`penanggung_jawab (x,y) → Akibat(x), Fasilitas(y)`

Gejala	Digunakan untuk melakukan ekstraksi bagian anggota tubuh yang terserang penyakit atau menderita kesakitan. `Gejala (x,y) → Populasi_penyakit(x,y)`
Obat	Digunakan untuk melakukan ekstraksi informasi obat atau vaksin, sebagai salah satu metode pemulihan `Obat (x,y) → Sebab_1(x), Pemulihan(y)`
Diagnosa	Digunakan untuk melakukan estraksi informasi yang terkait dengan pernyataan medis tentang penyakit `diagnosa (x,y) → Populasi_penyakit(x), gejala(y)`

III. PENGUJIAN

Berikut ini adalah salah satu contoh *dataset* yang digunakan dalam makalah ini[1].

> TEMANGGUNG - Penderita diare di Desa Sigedong, Tretep, Kabupaten Temanggung, Jawa Tengah yang dinyatakan sebagai kejadian luar biasa (KLB) jumlahnya terus bertambah.
>
> Kepala Dinas Kesehatan Kabupaten Temanggung, Suparjo mengatakan data hingga pagi tadi jumlah penderita mencapai 64 orang.
>
> Seperti diwartakan sebelumnya, pada Selasa (8/8) sore jumlah penderita diare di Desa Sigedong sebanyak 59 orang.
>
> Ia mengatakan ada seorang korban meninggal dunia dari kasus KLB tersebut. Korban sudah berusia lanjut 75 tahun, selain diare dia juga menderita hipertensi. Penyebab kematian apakah dari diare atau hipertensi belum diketahui.
>
> "Jumlah penderita terus bertambah, kapan berhenti kami tidak tahu, karena hingga saat ini penyebab KLB diare juga belum diketahui," katanya.
>
> Namun, katanya diduga kuat karena air yang dikonsumsi masyarakat dan saat ini masih dilakukan penelitian. Ia menuturkan dari sejumlah kasus tersebut terdapat lima penderita yang dirujuk untuk menjalani rawat inap, yakni dua pasien di RSK Parakan dan tiga pasien di Puskesmas Ngadirejo.
>
> Namun, kini mereka sudah pulang dari rumah sakit tersebut. Ia menuturkan untuk penanganan kasus KLB tersebut, Dinkes Temanggung telah mendirikan posko di desa tersebut yang buka 24 jam. "Kami buka posko selama 24 jam untuk penanganan penderita diare yang muncul dan jika mungkin perlu dirujuk kami juga siapkan ambulans di posko," ucapnya.
>
> Ia menuturkan guna mengantisipasi kasus tersebut, pihaknya juga melakukan sosialisasi kepada masyarakat untuk menerapkan hidup bersih dan sehat. Selain itu, juga dilakukan penebaran kaporit di mata air dan penampungan air untuk menurunkan angka bakteri dan kuman.

A. *Anotasi Frasa*

Pengujian diawali dengan mendapatkan anotasi frasa untuk mendapatkan makna negatif dan makna positif dari sebuah kalimat medis. Setiap kata pada sebuah kalimat diubah menjadi label *pos-tagging* dengan menggunakan *POS Tagset Penn Treebank*.

Kalimat awal:
```
Penyebab kematian apakah dari diare atau
hipertensi belum diketahui.
```

Proses *pos-tagging* menjadi:
```
Penyebab/(VB) kematian/(VB) apakah/(RB) dari/(CC)
diare/(NN) atau/(CC) hipertensi/(RB) belum/(RB)
diketahui/(VB) ./(Z)
```

Proses *pos-tagging* + anotasi *phrase* menjadi:
```
Penyebab/(VB) kematian/(VB + NEG) apakah/(RB)
dari/(CC) diare/(NN) atau/(CC) hipertensi/(RB)
belum/(RB) diketahui/(VB) ./(Z)
```

Hasil analisis:
```
Negatif
```

Anotasi frasa yang dijadikan acuan adalah ketika sebuah kata memiliki hasil *pos-tagging* VB maka makna dari kata tersebut adalah akibat. Ketika sebuah kata memiliki hasil *pos-tagging* NN, maka makna dari kata tersebut adalah sebab.

B. *Anotasi Paragraf*

Anotasi paragraf digunakan untuk menentukan makna paragraf yang terdapat pada sebuah artikel, termasuk ke dalam kategori paragraf inti, paragraf penjelas, atau paragraf kesimpulan. Tabel IV adalah hasil klasifikasi dari anotasi paragraf dengan menggunakan aturan yang terdapat pada Tabel II.

C. *Anotasi Elemen Medis*

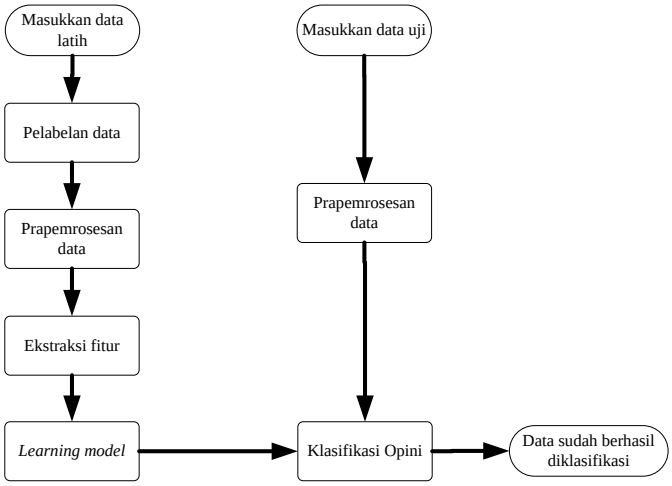
Anotasi ini merupakan gabungan dari anotasi frasa + anotasi paragraf + anotasi elemen medis. Dikarenakan penelitian ini merupakan *subjective classifier*, maka model

[1] https://lifestyle.okezone.com/read/2017/08/09/481/1752519/desa-sigedong-temanggung-klb-diare-1-korban-meninggal-dunia

pembelajaran harus disesuaikan berdasarkan preferensi individual pemberi label atau seorang pakar dan hasil label dapat berbeda-beda. Oleh karena itu, aktivitas penelitian melakukan eksplorasi fitur dengan cara klasifikasi seperti yang ditunjukkan pada Gbr. 5.

TABEL IV
HASIL KLASIFIKASI ANOTASI PARAGRAF

Kalimat medis	Anotasi Paragraf
(sistem) TEMANGGUNG$_{LOKASI}$ – Penderita PERSON diare PENYAKIT di Desa Sigedong, Tretep, Kabupaten Temanggung, Jawa Tengah LOKASI, yang dinyatakan sebagai kejadian luar biasa (KLB) jumlahnya terus bertambah.	Penjelas
(sistem) Kepala Dinas Kesehatan PERSON Kabupaten Temanggung LOKASI, Suparjo PERSON mengatakan data hingga pagi tadi jumlah penderita PERSON mencapai 64 NOMOR orang PERSON.	Penjelas
(sistem) Seperti diwartakan sebelumnya, pada Selasa (88) NOMOR sore jumlah penderita PERSON diare PENYAKIT di Desa Sigedong LOKASI sebanyak 59 NOMOR orang.	Penjelas
(sistem) Penyebab kematian AKIBAT apakah dari diare PENYAKIT atau hipertensi PENYAKIT belum diketahui.	Inti



Gbr. 5 *Subjective classifier*.

Metode yang digunakan untuk melakukan klasifikasi antara lain *naive Bayes* dan HMM. Berikut ini adalah *pseudocode* yang digunakan pada metode *naive Bayes* untuk melakukan klasifikasi anotasi elemen medis.

Anotasi menggunakan *naive Bayes*
a. Ambil data training dari basis data.
b. Looping berdasarkan banyaknya kalimat uji.
   1. Cari kemungkinan **TAG** dari masing-masing kata pada kalimat yang diuji di data training
   2. Jika terdapat kemungkinan **TAG**, buat unik **TAG** dari kemungkinan **TAG**. Jika tidak terdapat kemungkinan **TAG**, ambil unik **TAG** dari data training
      - Looping berdasarkan kata pada kalimat yang diuji
        - Menghitung matrix probabilitas antar kata uji dengan unik **TAG** ~ P(Xt|Yt)*P(Yt)
        - (1 + jumlah kata pada data training terhadap **TAG**) / jumlah **TAG** pada data training * (1 + jumlah **TAG** pada data training) / jumlah data training
      - End looping
   3. Mencari **TAG** dari tiap kata uji dengan mencari nilai terbesar dari hubungan antara kata uji dengan **TAG** yang ada pada matrix probabilitas
   4. Jika memilih Naïve Bayes dengan *"reinforcement"* simpan hasil **TAG** terpilih ke data training pada database. Jika hanya memilih Naïve Bayes maka hasil **TAG** terpilih tidak disimpan ke data training di database.
c. End looping
Menghitung *recall, precision & f-measure* dari hasil perhitungan Naïve Bayes

Sementara itu, *pseudocode* metode HMM untuk melakukan klasifikasi anotasi elemen medis adalah sebagai berikut.

Anotasi menggunakan *Hidden Markov Model*
a. Ambil data training dari database
b. Looping berdasarkan banyaknya kalimat uji
   1. Cari kemungkinan **TAG** dari masing-masing kata yang diuji di data training
   2. Jika terdapat kemungkinan **TAG**, buat unik **TAG** dari kemungkinan **TAG**. Jika tidak terdapat kemungkinan **TAG**, ambil unik **TAG** dari data training
   3. Membuat *"emission matrix"* atau matrix probabilitas kata pada kalimat uji terhadap **TAG** ~ P(Xt|Yt)
      - Looping berdasarkan kata uji
        - (1 + jumlah kata pada data training terhadap **TAG**) / jumlah **TAG** pada data training
      - End Looping
   4. Membuat *"transition matrix"* atau matrix probabilitas antar unik **TAG** ~ P(Yt|Yt-1)
      - Looping berdasarkan kata uji
        - (1 + jumlah kata pada data training terhadap **TAG**) / jumlah **TAG** pada data training
      - End Looping
   5. Membuat *"viterbi matrix"* dari *"emission matrix"* dan *"transition matrix"* yang telah dibuat
   6. Mencari **TAG** dari tiap kata uji dengan mencari nilai terbesar dari hubungan antara kata uji dengan **TAG** yang ada pada *"viterbi matrix"*

```
    7. Jika memilih HMM dengan "Reinforcement"
       simpan hasil TAG terpilih ke data
       training pada database. Jika hanya
       memilih HMM maka hasil TAG terpilih tidak
       disimpan ke data training di database.
c. End Looping
Menghitung recall, precision & f-measure dari
hasil perhitungan HMM
```

Hasil yang didapatkan pada anotasi elemen medis adalah sebagai berikut.

Kalimat awal:
`(sistem) Penyebab kematian apakah dari diare atau hipertensi belum diketahui.`

Proses *pos-tagging* menjadi:
`Penyebab/(VB) kematian/(VB + NEG)`
`apakah/(RB) dari/(CC) diare/(NN) atau/(CC) hipertensi/(RB) belum/(RB) diketahui/(VB) ./(Z)`

Penjelasan:
`Penyebab kematian` pada tahapan anotasi *phrase* termasuk ke dalam *pos-tagging* `VB+NEG` maka menjadi indikasi bahwa kata tersebut termasuk ke dalam kelas `akibat`.

Anotasi paragraf:
Hasil analisis pada anotasi paragraf adalah `paragraf inti`. Elemen yang terdapat pada kalimat tersebut antara lain (1) jumlah dalam angka: jumlah kejadian, bukti kejadian; (2) Akibat (3) Farmasi; (4) Diagnosis; (5) tanda di dalam tubuh terkait dengan gejala.

Sehingga hasil klasifikasi adalah
`(sistem) Penyebab kematian` AKIBAT `apakah dari diare` PENYAKIT `atau hipertensi` PENYAKIT `belum diketahui.`

Perbandingan hasil klasifikasi untuk anotasi frasa + anotasi paragraf + anotasi elemen medis dengan metode klasifikasi *naive Bayes* dan HMM seperti pada Tabel V.

TABEL V
PERBANDINGAN HASIL KLASIFIKASI ANOTASI FRASA + ANOTASI PARAGRAF + ANOTASI ELEMEN MEDIS

Metode	*Recall*	Presisi	*f-measure*
Naïve Bayes	0,905	0,924	0,910
HMM	0,706	0,750	0,720

TABEL VI
HASIL ANOTASI ELEMEN MEDIS KELUARAN SISTEM

Elemen medis	Hasil klasifikasi
1	**2**
LOKASI	Temanggung, desa Sigedong, Tretep, kabupaten Temanggung, Jawa Tengah, Parakan, Ngadirejo
PERSON	penderita, kepala dinas kesehatan, suparjo, orang, pasien
PENYAKIT	diare, hipertensi
NOMOR	64, 88, 59, lima, dua, tiga, 24
AKIBAT	**kematian**
PENGOBATAN	dirujuk, rawat inap, penanganan
FASILITAS	RSK, puskesmas, posko, ambulans
SEBAB	**bakteri, kuman**

Tabel VI merupakan hasil akhir yang diberikan oleh sistem. Elemen medis yang diusulkan mampu digunakan untuk melakukan identifikasi sebab-akibat dari artikel medis *online*.

IV. KESIMPULAN

Anotasi elemen medis dapat digunakan sebagai panduan untuk memberikan label pada artikel medis, sedangkan anotasi paragraf dapat digunakan untuk menentukan makna yang sedang dibahas pada paragraf tertentu. Gabungan kedua anotasi tersebut dapat digunakan untuk melakukan identifikasi hubungan sebab-akibat. Hasil evaluasi kinerja sistem dengan menggunakan gabungan anotasi frasa + anotasi paragraf + anotasi elemen medis dan metode klasifikasi *naive Bayes* mendapatkan hasil 0,905, 0,924, dan 0,910 secara berurutan untuk *recall*, presisi, dan *f-measure*. Namun, terdapat pemasalahan pada pemberian label secara *subjective classifier*. Keberhasilan kinerja sistem sangat dipengaruhi oleh seorang pakar pada domain tertentu untuk memberikan label tersebut. Usulan pada penelitian berikutnya adalah melakukan analisis pola anotasi frasa. Turunan anotasi frasa menjadi *semantic role labeling* dapat digunakan sebagai metode alternatif baru untuk melakukan identifikasi hubungan sebab-akibat yang tidak tergantung pada domain tertentu.

REFERENSI

[1] S. Akbar, L. Slaughter, dan Ø. Nytrø, "Collecting Health Related Text from Patient Health Writings," *The 2nd International Conference on Computer and Automation Engineering*, 2010, Vol. 1, hal. 15–19.

[2] X. Wu, X. Zhu, G.-Q. Wu, dan W. Ding, "Data Mining with Big Data," *IEEE Trans. Knowl. Data Eng.*, Vol. 26, No. 1, hal. 97–107, 2014.

[3] A.J. Yepes, A. MacKinlay, B. Han, dan Q. Chen, "Identifying Diseases, Drugs, and Symptoms in Twitter," *Stud. Health Technol. Inform.*, Vol. 216, hal. 643–647, 2015.

[4] A.J. Yepes, A. MacKinlay, dan B. Han, "Investigating Public Health Surveillance using Twitter," *Proc. 2015 Work. Biomed. Nat. Lang. Process. (BioNLP 2015)*, 2015, hal. 164–170.

[5] K. Byrd, A. Mansurov, dan O. Baysal, "Mining Twitter Data for Influenza Detection and Surveillance," *Proc. Int. Work. Softw. Eng. Healthc. Syst. - SEHS '16*, 2016, hal. 43–49.

[6] C.D. Corley, D.J. Cook, A.R. Mikler, dan K.P. Singh, "Text and Structural Data Mining of Influenza Mentions in Web and Social Media," *Int. J. Environ. Res. Public Health*, Vol. 7, No. 2, hal. 596–615, 2010.

[7] L. Wu, T.-S. Moh, dan N. Khuri, "Twitter Opinion Mining for Adverse Drug Reactions," *2015 IEEE Int. Conf. Big Data (Big Data)*, 2015, hal. 1570–1574.

[8] Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller, R.M. Massanari, "A Potential Causal Association Mining Algorithm For Screening Adverse Drug Reactions In Postmarketing Surveillance," *IEEE Trans. Inf. Technol. Biomed.*, Vol. 15, No. 3, hal. 428–437, 2011.

[9] J. Atkinson dan A. Rivas, "Discovering Novel Causal Patterns from Biomedical Natural-Language Texts Using Bayesian Nets," *IEEE Trans. Inf. Technol. Biomed.*, Vol. 12, No. 6, hal. 714–722, 2008.

[10] S.B. Bhaskoro, S. Akbar, dan S.H. Supangkat, "Identification of Causal Pattern using Opinion Analysis in Indonesian Medical Texts," *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2015, hal. 1–7.