Jurnal INFORMA Politeknik Indonusa Surakarta p-ISSN : 2442-7942, e-ISSN 2716-5051
Vol. 6 Nomor 1 Juni 2020

Membangun Basis Pengetahuan Untuk Interaksi Obat Dengan Sosial Media

Nanang Prasetiyantara[1], Kusrini[2], Asro Nasiri[3]

[1,2,3]Magister Teknik Informatika, Universitas AMIKOM Yogyakarta
Jl. Ringroad Utara, Condong Catur, Sleman, Yogyakarta 55283 Indonesia
Email : [1]prasetiyantara@gmail.com, [2]kusrini@amikom.ac.id, [3]asro@amikom.ac.id

Abstract

With many adults using social media to discuss health information, researchers have begun to dive into this resource to monitor or detect health conditions at the population level. Twitter, in particular, has grown to several hundred million users and can attend rich source of information for detecting serious medical conditions, such as adverse drug reactions (ADRs). However, Twitter too presents unique challenges due to brevity, lack of structure, and informal language. We crawled data from Twitter presenting 10,822 freely available tweets, which can be used to train automated tools to mine Twitter for ADR. We collect tweets using drug names as keywords, but expanding it by applying the Natural Language Processing (NLP) algorithm to produce misspelled versions of drug names for and drug interactions. We annotate each tweet for the presence of mentioning interactions, and for those who have, mention annotations. Agreement between our annotators for binary classification. We evaluate the usefulness of the dataset with machine learning algorithm training classes: using C.45.

Keywords:*Natural Language Processing (NLP), C.4, Twitter*

1. PENDAHULUAN

Obat merupakan bahan atau paduan bahan, termasuk produk biologi yang digunakan untuk mempengaruhi atau menyelidiki system fisiologi atau keadaan patologi dalam penetapan diagnosis, pencegahan, penyembuhan, pemulihan serta peningkatan kesehatan pada kesehatan manusia. Bahan baku obat yang berkhasiat maupun tidak dalam pengolahannya terdapat standard an mutu bahan baku farmasi. (Menkes RI, 2013)

Reaksi obat dapat berbentuk berbagai macam dari yang berbahaya hingga menyenangkan akibat penggunaan obat tertentu. Dikemudian hari hal ini dapat menjadi masalah terutama terkait administrasi jenis obat tertentu, dosis penggunaan obat bahkan penarikan produk (Yang & Christopher C. Yang, 2015; Christopher C. Yang, 2016).

Kehadiran media sosial menjadi banyak perhatian bagi sebagian orang. *Facebook, Instagram, Youtube, Twitter*, beberapa media sosial yang diminati banyak orang. Peneliti ingin mengetahui interaksi obat yang dirasakan oleh konsumen melalui media sosial dengan pengolahan Bahasa *Natural Language Processing (NLP)* agar lebih tertata strutukur bahasanya dan kemudian di klasifikasikan dengan C.45.

2. METODE PENELITIAN

a. Jenis Penelitian

Jenis penelitian yang digunakan adalah penelitian kuantitatif dimana penelitian kuantitatif menjadi faktor penting dalam proses penelitian itu sendiri. Bahwa sebagaian dari kegiatan penelitian adalah proses teori atau berteori. Pada proses penelitian ini melakukan proses analisis deduktif untuk menguji atas permasalahan yang sedang dihadapi. Pada penelitian kuantitatif baik teori ataupun acuan teori digunakan untuk membantu peneliti menemukan masalah penelitian, menemukan hipotesism menemukan konsep, menemukan metodologi dan menemukan *tools* analisis data

b. Pendekatan Penelitian

Pada penelitian ini menggunakan pendekatan penilitian kuantitatif dimana data yang digunakan adalah data dari sosial media twitter.

c. Pendekatan Penelitian

Dalam penelitian ini, pengumpulan data yang akan digunakan menggunakan beberapa langkah yang berkaitan dengan metode penelitian tersebut, yaitu dengan metode crawling data media sosial dan studi kepustakaan.

1) Crawling Data

Crawling data adalah metode pengumulan data melalui media sosial secara terstruktur dengan etika tertentu. Aplikasi yang menjalanakan proses crawling disebut web crawler.

2) Studi Kepustakaan

Tinjauan pustaka dilakukan untuk mengetahui posisi teoritis yang akan dikembangkan. Peneliti melakukan pencarian literature dan pengumpulan materi berkaitan dari jurnal, buku, dan referensi yang lainnya.

d. Metode Analisis Data

Pada tahapan penelitian ini akan dilakukan proses analisis data dari data yang telah diperoleh sebagai bahan penelitian. Analisis data merupakan proses penyederhanaan data dalam bentuk yang lebih mudah dibaca dan diimplementasikan. Analisa data diperlukan untuk mengetahui seberapa besar tingkat akurasi data yang diperoleh dari subyek penelitian

e. C.45

Dalam proses pengujian attribut, cabang baru yang terbentuk akan diperhatikan dari tipe attribut (Jiawei Han, 2012) terdapat 3 jenis cabang yang mungkin muncul dalam pohon keputusan, yaitu:

1) Jika attribut bernilai diskrit, maka cabang yang terbentuk akan selalu sama dengan jumlah variasi nilai yang terdapat pada attribut tersebut.
2) Jika cabang bernilai kontinyu, maka akan dipecahkan menurut titik perpecahan, sedangkan titik perpecahan dikalkulasi dengan masing masing algoritma penyusun pohon keputusan. Cabang perpecahan yang terbentuk akan berpola seperti ≤ attribut, dan satu cabang lagi > attribute
3) Jika attribut yang diuji bernilai biner, maka cabang yang terbentuk pasti dua dan melibatkan nilai ya atau tidak

Perhitungan alogaritma C.45 sebagai berikut:

1) Perhitungan *Entropy* dan *Gain*

Langkah pertama menghitung Entropy, kemudian nilai Entropy seluruh nilai fitur per atribut dengan rumus :

$$Entropy(X) = \sum_{j=1}^{k} -pj \times \log 2\, pj$$

Keterangan:
X : Himpunan kasus
K : Jumlah partisi X
Pj: Proporsi Xj terhadap X

Kemudian menghitung nilai Gain setiap atribut dengan rumus

$$Gain(X, A) = Entropy(X) - \sum_{j=1}^{k} \frac{|X_i|}{|X|} * Entropy(X_i)$$

Keterangan:
X: Himpunan kasus
A: Atribut
Xi: Proporsi atribut ke X terhadap jumlah kasus

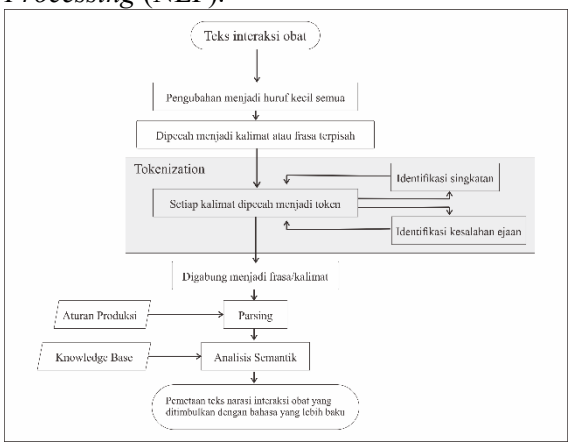
2) Pembentukan pohon keputusan

Pembentukan pohon dimulai dari root dengan memilih atribut yang memiliki Gain tertinggi.

f. *Natural Language Processing* (NLP)

Bahasa alami (natural language) adalah bahasa yang biasa digunakan oleh manusia untuk berkomunikasi12. Istilah "Natural Language Processing" (NLP) biasanya digunakan untuk mendeskripsikan fungsi dari komponen perangkat lunak atau perangkat keras pada sistem komputer yang dapat menganalisis atau menyintesis bahasa alami, baik lisan ataupun tulisan (teks)[4]

Berikut bagan proses *Natural Language Processing* (NLP).



Penjelasa dari gambar 1 adalah sebagi berikut:

a. Teks interaksi obat

Teks interaksi obat ini dapat berupa satu kalimat atau lebih. Kalimat tersebut dapat

terdiri dari satu frasa/klausa atau lebih. Apabila dalam satu kalimat terdapat lebih dari satu frasa/klausa, maka dipisahkan dengan koma

b. Pengubahan menjadi huruf kecil

Masukan berupa teks interaksi obat diubah menjadi huruf kecil semua untuk nantinya mempermudah dalam proses-proses selanjutnya.

c. Pemecahan menjadi frasa terpisah

Pemecahan kalimat dilakukan berdasarkan karakter titik ataupun new line. Sedangkan pemecahan frasa/klausa dilakukan berdasarkan karakter koma.

d. Tokenization

Kalimat-kalimat dan frasa-frasa/klausa-klausayang telah diperoleh dari tahap sebelumnya, dilakukan pemecahan menjadi token-token (tokenization), dengan menggunakan karakter spasi atau tanda baca sebagai pemisah Dalam tahap ini dilakukan exception apabila terdapat karakter titik, koma, atau garis miring yang diapit oleh karakter angka, dalam hal ini tidak dilakukan pemecahan token. Dari token-token yang diperoleh, dilakukan identifikasi singkatan, mengingat pencatatan anamnesis interaksi obat yang ditulis di media sosial tidak terlepas dari penggunaan singkatan yang tidak baku. Sebagai contoh, kata “sll” untuk “selalu”. Pendekatan yang digunakan untuk identifikasi singkatan tersebut adalah menggunakan leksikon / kamus. Setelah itu dilakukan identifikasi kesalahan ejaan. Setiap token akan dicari pada leksikon, jika tidak ditemukan akan dianggap sebagai kata yang salah eja, sehingga akan dilakukan spelling correction. Pendekatan yang digunakan adalah edit distance, yaitu dengan melakukan perhitungan kemiripan antara token yang tidak ditemukan dalam leksikon tadi dengan kata-kata yang terdapat dalam leksikon.

e. Penggabungan kembali menjadi kalimat

Token yang sudah dilakukan identifikasi singkatan dan bebas dari kesalahan ejaan akan digabung kembali menjadi frasa/klausa dan/atau kalimat terpisah.

f. Parsing

Berdasarkan data yang diperoleh dari dokter, dibuat aturan produksi untuk memetakan kalimat masukan keluhan pasien ke dalam objek-objek riwayat penyakit sekarang, yang nantinya dapat membentuk bahasa baku/bahasa medis yang tepat dari serangkaian pemetaan objek yang terbentuk

g. Pemetaan Interaksi obat di media sosial

Langkah terakhir setelah parsing dengan mengabungkan menjadi satu buah kalimat yang memilik makna.

3. TINJAUAN PUSTAKA

Penelitian Corinna Kola´rˇik (2007) mengidentifikasi istilah obat-obatan secara otomatis dengan metode pendekatan pola *Hearst* untuk ekstrasi istilah dari DrugBank secara gratis untuk menggambarkan property obat yang terkait langsung. Dengan demikan dapat ditemukan anotasi obat baru yang belum ada dalam database obat sekaligus menambah anotasi informasi anotasi obat baru.

Penelitian dari Elise Bigeard (2018) dengan menganalisis penyalahgunaan obat yang dilakukan oleh pasien berdasarkan informasi dari sosial media dengan istilah seperti depresi (depresi), *anxiété* (kecemasan), *nervux* (gugup), *fobie* (fobia), *panique* (panik), atau *angoisse* (kesulitan). Pendekatan menggunakan *Lexique.org*, Wikidata dan JeuxdeMots dengan alogaritma distribusi. Menghasilkan topologi penyalahgunaan obat dalam pesan di forum kesehatan Perancis dengan situs web Doctissimo, hingga 60% membicarkan pengendalian kelahiran, 15 % membicarkan tentang antidepresan dan *anxiolytics*. Sedangkan di forum kehamilan hanya 44% membahas terkait 3 kelas obat.

4. HASIL DAN PEMBAHASAN

Data yang diambil pada penelitian ini hasil dari crawling data dari sosial media twitter yang berjumlah sekitar 300 data, kemudian dilakukan preprocessing dengan melakukan pelabelan seperti berikut:

Tabel 1. Tabel Label

No	Nama Pelajaran
1	Batuk
2	Flu
3	Sakit Mata

Sedangkan untuk atributnya sendiri sebagai berikut

Tabel 2. Tabel Atribut

No	Nama Pelajaran
1	Batuk
2	Flu
3	Sakit Mata

Tabel 3. Hasil Proses Training dan Testing

Dataset	**Sumber Data**	**Jml Data**	***Acc* (%)**	***Prediksi***
Data *training*	Data Twitter	200	100	
		200	89,6	
Data *testing*	Data Excel	50		75

Hasil perhitungan akhir entropy dan gain dan pohon keputusan masih ada nilai atribut yang belum memasuki kelas. Tetapi karenahanya atribut lama efek yang digunakan maka tidak ada perhitungan lagi untuk nilai atribut yang belum memasuki kelas. Maka perhitungan selesai sampai di sini

		Jumlah	Ngantuk	Pusing	Enthropy
Total		8	5	3	
Jenis Penyakit	Batuk	4	5	3	-2,96656
	Flu	4	2	2	-0,08048

Gambar 1. Perhitungan Entrophy

5. KESIMPULAN

Berdasarkan pembahasan yang sudah dilakukan menggunakan C.45 dan evaluasi performance dari 10 sample data twitter menunjukan prediksi 75 dari 100 maka hal ini ada pengetahuan baru sebesar 25 hal ini cukup baik, dengan 5 atribut yang digunakan. Untuk selanjutnya disarankan menggunakan algoritma yang berbeda dengan data yang sama sehingga bisa membandingkan dari hasil yang sudah dilakukan.

6. REFERENSI

Christopher C. Yang, L. J. (2016). Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media. *Proceedings of ACM SIGKDD Workshop on Health Informatics*.

Elise Bigeard. (2018) Detection and Analysis of Drug Misuses. A Study Based on Social Media Messages, *Frontiers in Pharmacology*

Corinna Kola´ rˇik. (2007) Identification of New Drug Classification Terms in Textual Resources, Bioinformatics

Chanifah Indah Ratnasari, Sri Kusumadewi, Linda Rosita (2014) Model Natural Language Processing untuk Perumusan Keluhan Pasien, *Seminar Nasional Informatika Medis (SNIMed) V 2014.*