REPOSITOR, Vol. 2, No. 2, Februari 2020, Pp. 193-200
ISSN : 2714-7975
E-ISSN : 2716-1382

Ekstraksi Informasi Kesehatan Masyarakat Dari Tweet Berbahasa Indonesia Berbasis Klasifikasi Dengan Algoritma Naive Bayes

Khoirir Rosikin*[1], Setio Basuki[2], Yufis Azhar[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
khoirir.rosi24@gmail.com[*1], setiobasuki.umm@gmail.com[2,] , yufis@umm.ac.id[3]

Abstrak

Kesehatan merupakan kebutuhan utama manusia. Di Indonesia terdapat permasalahan tentang kesehatan, yaitu meningkatnya penyakit menular dan penyakit tidak menular. Untuk mengatasinya perlu dilakukan tidakan pencegahan. Salah satu usaha untuk melakukan pencegahan penyakit, adalah dengan mengetahui informasi penyakit tersebut, temasuk tentang penyebab dan akibat yang ditimbulkan, sehingga bisa melakukan pencegahan. Informasi bisa didapatkan dengan berbagai macam cara, salah satunya diambil dari media sosial, terutama twitter. Twitter digunakan karena banyaknya tweet yang dihasilkan sehingga memunculkan fenomena big data. Karena hal itulah, penelitian ini bermaksud untuk melakukan suatu metode ekstraksi informasi. Ekstraksi informasi merupakan metode penerapan data mining terutama bidang text mining yang digunakan untuk mendapatkan informasi dari kumpulan banyak data. Informasi yang dimaksud adalah penyakit, akibat, dan penyebab. Penelitian ini menggunakan pendekatan ekstraksi informasi berbasis klasifikasi dengan algoritma Naive Bayes. Penelitian ini menggunakan 7 set fitur dan sebuah model algoritma klasifikasi yaitu Naive Bayes. Dalam ekstraksi fitur terjadi imbalance dataset, sehingga dilakukan resample filtering data. Pengujian dilakukan dengan 2 metode, yaitu pengujian model dengan menggunakan 10-folds cross-validation dan pengujian klasifikasi dengan menggunakan 100 data uji. Hasil dari pengujian model mendapatkan nilai akurasi 77,27% dan pengujian klasifikasi mendapatkan nilai akurasi 74,07%.

Kata Kunci: *Ekstraksi Informasi, Klasifikasi, Naive Bayes, NLP*

Abstract

Health is a primary human need. In Indonesia there are health problems, namely the increase of infectious diseases and non-communicable diseases. To overcome this need to do precautionary measures. One effort to prevent disease, is to know the disease information, including about the causes and effects caused, so it can do prevention. Information can be obtained in various ways, one of which is taken from social media, especially twitter. Twitter is used because of the number of tweets produced resulting in big data phenomenon. Because of that, this research intends to perform an information extraction method. Information extraction is a method of application of data mining, especially the text mining field used to obtain information from a large collection of data. The information in question is a disease, effect, and cause. This research uses a classification-based information extraction approach with Naive Bayes algorithm. This research uses 7 feature sets and a model of classification algorithm that is Naive Bayes. In feature extraction there is imbalance dataset, so it is done resample filtering data. The test is done by 2 methods, namely model testing using 10-folds cross-validation and classification testing using 100 test data. The result of model test get the accuracy value 77,27% and the classification test get the accuracy value 74,07%.

Keywords: *Information Extraction, Classification, Naive Bayes, Natural Language Processing*

1. Pendahuluan

Kesehatan merupakan kebutuhan hidup manusia yang paling utama. Menurut UU kesehatan nomor 36 TAHUN 2009, kesehatan merupakan suatu kondisi yang sehat, secara fisik, mental, rohani, dan juga secara sosial yang memungkinkan setiap orang bisa tetap hidup baik secara sosial maupun ekonomis. Menurut [1], terdapat 6 permasalahan kesehatan yang ada di Indonesia, salah satunya adalah meningkatnya penyakit menular dan penyakit tidak menular. Hal ini tentu membuat menurunnya kesehatan masyarakat di Indonesia. Menurut Winslow [2],

kesehatan masyarakat (*public health*) merupakan suatu ilmu atau seni yang berkaitan dengan pencegahan penyakit melalui beberapa usaha. Salah satu usaha untuk melakukan pencegahan penyakit, adalah dengan mengetahui informasi penyakit tersebut, temasuk tentang penyebab dan akibat yang ditimbulkan, sehingga bisa melakukan pencegahan [2].

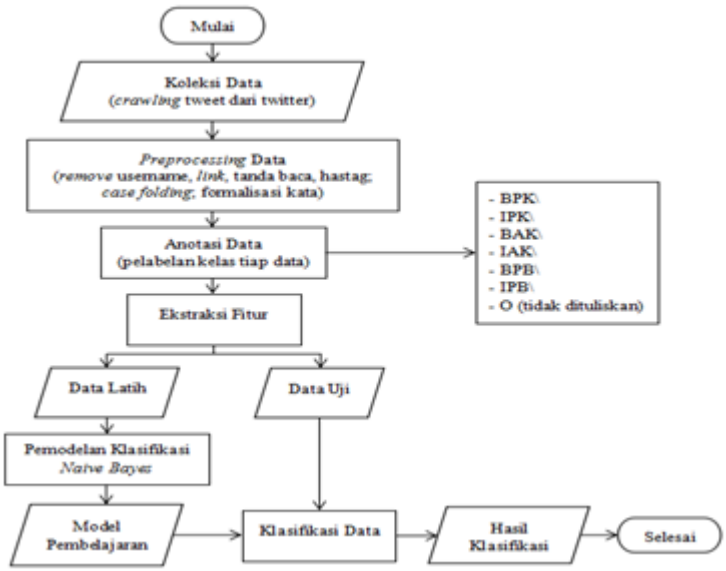
Informasi merupakan hasil dari pengolahan data yang bernilai guna rendah menjadi bernilai guna tinggi bagi penerimanya [3]. Salah satu penghasil data terbesar adalah media sosial, salah satunya twitter. Pengguna twitter di seluruh dunia pada tahun 2017 mencapai 320 juta, sedangkan jumlah pengguna twitter di Indonesia jumlahnya belum bisa dipastikan, tetapi termasuk 5 besar pengguna terbanyak twitter. Menurut data yang dipublikasikan oleh Twitter Indonesia pada tahun 2016, menyebutkan bahwa sekitar 77% pengguna twitter di Indonesia adalah pengguna aktif dan menghasilkan sekitar 4,1 miliar tweet sepanjang tahun 2016. Menurut [4] menyebutkan semakin besarnya pertumbuhan pengguna twitter, semakin banyak juga data yang dihasilkan, fenomena ini disebut dengan *big data*. *Big data* bisa digunakan untuk mendapatkan informasi yang bisa dimanfaatkan [5], termasuk mencari informasi tentang penyebab dan akibat dari suatu penyakit. Dalam penerapan *data mining* terutama bidang *text mining*, untuk mendapatkan informasi dari kumpulan banyak data dapat dilakukan dengan metode ekstraksi informasi [5].

Ekstraksi informasi pada twitter dimanfaatkan untuk mengetahui komplain masyarakat terhadap pemerintahan kota Bandung [5]. Peneliti [6] [7] melakukan ekstraksi informasi terhadap twitter untuk mengumpulkan secara otomatis informasi transaksi online di Indonesia. Berdasarkan hal tersebut, penelitian ini menggunakan twitter sebagai data penelitian dengan memanfaatkan penerapan *text mining,* yaitu ekstraksi informasi berbasis klasifikasi. Informasi target yang dimaksud adalah kesehatan masyarakat (*public health*) terutama usaha pencegahan penyakit, yaitu mengetahui informasi Penyakit, Akibat, dan Penyebab. Berdasarkan referensi jurnal, penelitian ini menggunakan algoritma yang populer yaitu *Naive Bayes*. Hasil dari ekstraksi informasi berbentuk sebuah kata yang diklasifikasikan berdasarkan informasi target.

2. Metode Penelitian

2.1 Diagram Alir Eksperimen Sistem

Diagram alir penelitian ini menjelaskan alur kumpulan data yang diproses sampai kumpulan data tersebut diklasifikasi. Gambar 1 berikut merupakan diagram alir eksperimen system.



Gambar 1. Diagram Alir Eksperimen Sistem

Gambar 1 menjelaskan proses eksperimen penelitian. Koleksi data dilakukan *preprocessing* yang meliputi penghapusan *username*, *url*, dan tanda baca lalu dilakukan *case folding*, dan terakhir dilakukan formalisasi kata pada tweet. Setelah *preprocessing*, selanjutnya dilakukan anotasi secara manual oleh anotator, lalu data tersebut diekstraksi fitur. Setelah

diekstraksi fitur, data dibagi menjadi 2, yaitu data latih dan data uji. Data latih digunakan pada pelatihan sistem dan menghasilkan model pembelajaran dengan algoritma *Naive Bayes*. Data uji digunakan pada pengujian sistem yang kemudian diklasifikasikan berdasarkan *class* target.

2.2 Data Penelitian

Data penelitian didapatkan dengan melakukan crawling dari twitter dari tanggal 1 Januari 2017 s/d 1 Oktober 2017. Kata kunci yang digunakan untuk crawling, antara lain: batuk, pilek, flu, demam, dbd, cacar air, tipes, tifus, dan hepatitis. Hasil crawling disimpan dalam file dengan format *Comma Separated Value* (CSV).

2.3 *Preprocessing* Data

Data penelitian yang didapatkan dari crawling twitter, terlebih dahulu dilakukan *preprocessing* data. *Preprocessing* data dilakukan agar data menjadi valid, sehingga data bisa diproses oleh sistem. *Preprocessing* data yang digunakan pada penelitian ini antara lain, *remove username*, *remove link*, *remove hastag*, *remove punctuation*, *case folding*, dan formalisasi kata.

2.4 Anotasi Data

Tahap anotasi data merupakan tahap pelabelan manual pada data tweet. Anotasi dilakukan dengan menambahkan label kelas pada kata dari tweet yang mengandung informasi tentang: penyakit (PK), akibat (AK), dan penyebab (PB). Anotasi dilakukan pada tweet yang telah dilakukan preprocessing data. Anotasi data dilakukan oleh 2 orang *anotator* + 1 orang *validator*. *Validator* ditugaskan untuk memvalidasi anotasi dari masing-masing *anotator*, sehingga mendapatkan anotasi data yang valid.

Berdasarkan informasi yang akan diekstraksi dari tweet, tidak semua informasi merupakan sebuah kata yang berdiri sendiri, terkadang informasinya merupakan sebuah frase atau lebih dari satu kata yang mengandung informasi, sehingga diperlukan label tambahan yang berfungsi untuk mengenali frase serta bagian-bagiannya. Label tambahan yang dimaksud adalah label BIO (*Begin-In-Other*). Pada label BIO, setiap kata akan diberikan label berupa Begin, In, atau Other. Label *Begin* digunakan untuk menandai kata yang menjadi awal frase, label *In* digunakan untuk kata yang merupakan bagian dari frase di depannya (yang ber-label *Begin*), kemudian *Other*, merupakan label untuk kata yang tidak mengandung informasi. Untuk label *Other*, tidak dituliskan pada saat anotasi data, karena label Begin dan In sudah cukup mewakili frase-frase yang dibutuhkan. Tabel 1 berikut merupakan label yang digunakan untuk anotasi data.

Tabel 1. Label Anotasi Data

Label	Deskripsi
/BPK	Label untuk kata pertama yang mengandung informasi tentang penyakit.
/IAK	Label untuk kata kedua dan berikutnya yang mengandung informasi tentang penyakit.
/BAK	Label untuk kata pertama yang mengandung informasi tentang akibat.
/IAK	Label untuk kata kedua dan berikutnya yang mengandung informasi tentang akibat.
/BPB	Label untuk kata pertama yang mengandung informasi tentang penyebab.
/IPB	Label untuk kata kedua dan berikutnya yang mengandung informasi tentang penyebab.
OTHER	Label untuk kata yang tidak mengandung informasi target. (Label Other tidak dituliskan pada anotasi data)

Anotasi dilakukan agar tiap kata pada tweet memiliki fitur *class*. Sehingga tiap kata tersebut dapat diklasifikasikan berdasarkan *class* yang dianotasikan. Data yang telah dianotasi, kemudian diekstraksi fitur.

2.5 Ekstraksi Fitur

Ekstraksi fitur merupakan tahap yang difungsikan untuk mendapatkan ciri-ciri dari masing-masing kata untuk dijadikan *token*. Fitur difungsikan agar sistem mengenali ciri-ciri dari setiap kata termasuk ke dalam informasi Penyakit, Akibat, dan Penyebab. Fitur-fitur yang digunakan untuk mengenali setiap kata, didapatkan dengan mengacu pada penelitian-penelitian sejenis

sebelumnya dan juga jurnal referensi, serta tambahan atau rekayasa fitur, untuk meningkatkan kinerja sistem. Tabel 2 berikut merupakan fitur yang digunakan penelitian ini.

Tabel 2. Set Fitur

Fitur	Deskripsi
Token	Kata hasil tokenisasi yang diproses.
TipeToken	Tipe dari *token* yang diproses (NUM, WORD, PUNCT).
POS	*Part-Of-Speech*, (NN, VB, JJ, CDP, dll.) dari *token* yang diproses.
NE	*Name Entity* dari token yang diproses (PERSON, ORGANIZATION, LOCATION, QUANTITY, dll).
Bef1POS	*Part-Of-Speech* dari satu *token* sebelum *token* yang diproses.
Bef1NE	*Name Entity* dari satu *token* sebelum *token* yang diproses.
Class	Kelas dari *token* yang diproses.

Ekstraksi fitur akan dilakukan dengan menggunakan bahasa pemrograman Java, serta menggunakan *library* InaNLP.jar. Library InaNLP digunakan untuk mendapatkan fitur-fitur NLP, diantaranya tipe *token*, *Part-Of-Speech*, serta *Name Entity* dari setiap *token*.

Hasil ekstraksi fitur, sering kali terjadi kondisi di mana ukuran sebuah target kelas jauh lebih besar dibandingkan dengan kelas yang lain. Dalam *Data Mining*, hal tersebut dinamakan dengan *Imbalance Dataset*, yang dapat menyebabkan hasil perhitungan kinerja sebuah klasifikasi menjadi tidak akurat. Metode *sampling*, adalah sebuah metode yang cukup populer dalam menangani permasalahan *imbalance dataset* [8]. Pada penelitian ini, penulis mencoba mengaplikasikan filter *resample* dari WEKA untuk menangani *imbalance dataset* yang mungkin terjadi.

2.6 Pembentukan Model

Proses pembentukan model, data latih yang sudah terbentuk, akan diproses menjadi sebuah model dengan cara mengaplikasikan sebuah algoritma pada sebuah dataset. Hal ini dilakukan dengan menggunakan Java serta *library* WEKA *machine learning*, dan menggunakan algoritma yang sudah ditentukan sebelumnya, yaitu *Naïve Bayes*. Tahap ini merupakan tahap pelatihan, dimana program akan menghasilkan sebuah model yang dibentuk dari data latih serta algoritma yang digunakan.

2.7 Klasifikasi *Naive Bayes*

Naive Bayes merupakan sebuah metode atau algoritma klasifikasi yang mengadopsi teorema *Bayes*. *Naive Bayes* memberikan metode langsung untuk menghitung probabilitas tersebut. Lebih tepatnya, teorema *Bayes* memberi kemudahan menghitung probabilitas sebuah hipotesis berdasarkan probabilitas sebelumnya, probabilitasnya mengamati berbagai data yang diberikan hipotesis, dan data yang teramati itu sendiri. Pada teorema *Bayes*, bila terdapat dua kasus terpisah (misalkan *D* dan *h*), maka teorema *Bayes* dirumuskan sebagai Persamaan 1 [9].

$$P(h \mid D) = \frac{P(D \mid h).P(h)}{P(D)} \quad (1)$$

Proses klasifikasi *Naive Bayes* dilakukan dengan mengambil sebuah contoh data tweet, yang kemudian sistem akan mengklasifikasikan tiap kata dari data uji yang mengandung informasi target.

2.8 Pengujian

Pengujian penelitian menggunakan 2 metode pengujian, yaitu pengujian model dan pengujian klasifikasi. Pengujian model dilakukan menggunakan metode *k-folds Cross-Validation* dan dihitung akurasinya. Pengujian klasifikasi, menggunakan 100 data tweet sebagai data uji. Data uji dilakukan formalisasi kata dan data yang diujikan diklasifikasikan berdasarkan informasi target. Untuk mengetahui keberhasilan sistem klasifikasi, hasil dari klasifikasi sistem kemudian dibandingan dengan hasil klasifikasi manual dan dihitung akurasinya. Perhitungan hasil akurasi menggunakan *Confusion Matrix*. *Confusion Matrix* merupakan teknik yang digunakan untuk mengukur tingkat kebenaran atau keberhasilan dari proses klasifikasi [10].

3. Hasil Penelitian dan Pembahasan

3.1 Data Penelitian

Data penelitian didapatkan dengan melakukan crawling dari twitter. Data yang dicrawling merupakan tweet dari tanggal 1 Januari 2017 s/d 1 Oktober 2017. Kata kunci yang digunakan untuk crawling twitter, antara lain: batuk, pilek, flu, demam, dbd, cacar air, tipes, tifus, dan hepatitis. Hasil crawling mendapatkan data tweet sebanyak 1000 tweet dan data tersebut disimpan dalam file dengan format *Comma Separated Value* (CSV). Gambar 2 berikut merupakan sebagian dari data yang telah dikumpulkan.

@larrybenardo, Gara" pilek sama cuaca mainstream gini badan agak ngedrop . Ayo semangat
@halimahnurb1, Pilek gara gara tadi kehujanan.
@ichall05, Gak ... Bisa tidur ... Gara" pilek
@randysutrisna, Gabisa tidur gara" pilek kampret
@resna, pilek gara-gara kebanyakan makan mecin. semua memang salah mecin.
@Grafienaliefya, Susah napas gara pilek
@jnhjlw_, Gara gara pilek suara jd kayak tikus kejepit
@YuyunAinie, Hadeuh cape gara-gara pilek bersin2 mulu.

Gambar 2. Data Penelitian

Gambar 2 merupakan sebagian dari kumpulan data penelitian yang disimpan ke dalam file format CSV. Kumpulan data tersebut dilakukan *preprocessing* data untuk melakukan pembersihan terhadap data penelitian.

3.2 *Preprocessing* Data

Preprocessing data merupakan tahap pembersihan data, agar data menjadi valid, sehingga data bisa diproses oleh sistem. *Preprocessing* data yang digunakan pada penelitian ini antara lain: *remove username, remove link, remove hastag, remove punctuation, case folding*, dan formalisasi kata. Gambar 3 berikut merupakan hasil *preprocessing* data.

gara pilek sama cuaca mainstream begini badan agak ngedrop ayo semangat
pilek gara gara tadi kehujanan
tidak bisa tidur gara pilek
tidak bisa tidur gara pilek sialan
pilek gara gara kebanyakan makan mecin semua memang salah mecin
susah napas gara pilek
gara gara pilek suara jadi kayak tikus kejepit
hadeuh capek gara gara pilek bersin-bersin melulu

Gambar 3. Hasil Preprocessing Data

Data yang telah dilakukan *preprocessing* masih terdapat kekurangan, yaitu tanda baca pisah (*dash*) dan ada beberapa *link* URL yang belum bersih walaupun telah dilakukan pembersihan. Sehingga setelah *preprocessing* data selesai, masih ada satu proses yang dilakukan secara manual yaitu melakukan cek hasil *preprocessing* serta menghapus tanda baca dan beberapa *link* URL yang dapat mempengaruhi hasil klasifikasi.

3.4 Anotasi Data

gara BPK\pilek sama cuaca mainstream begini BAK\badan IAK\agak IAK\ngedrop ayo semangat
BPK\pilek BPB\gara IPB\gara IPB\tadi IPB\kehujanan
BAK\tidak IAK\bisa IAK\tidur gara BPK\pilek
BAK\tidak IAK\bisa IAK\tidur gara BPK\pilek sialan
BPK\pilek BPB\gara IPB\gara IPB\kebanyakan IPB\makan IPB\mecin semua memang salah mecin
BAK\susah IAK\napas gara BPK\pilek
gara gara BPK\pilek BAK\suara IAK\jadi IAK\kayak IAK\tikus IAK\kejepit
hadeuh capek gara gara BPK\pilek BAK\bersin IAK\bersin IAK\melulu

Gambar 4. Hasil Anotasi Data

Anotasi data merupakan tahap pelabelan manual pada data tweet yang telah dilakukan preprocessing. Anotasi dilakukan dengan menambahkan label kelas pada kata dari tweet yang

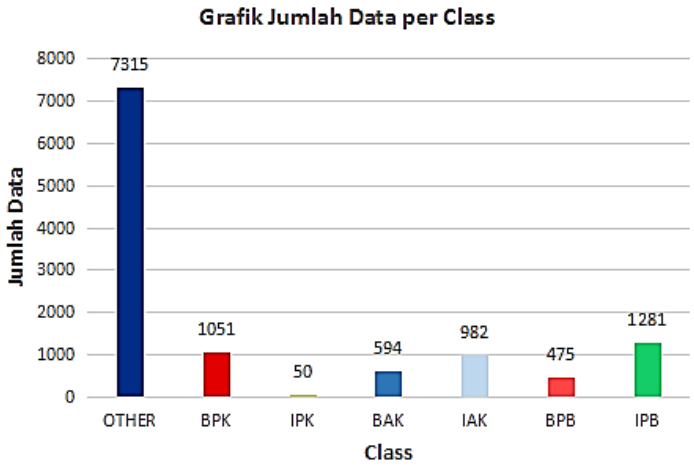
mengandung informasi tentang: penyakit (BPK dan IPK), akibat (BAK dan IAK), dan penyebab (BPB dan IPB). Fungsi anotasi adalah agar sistem mengetahui *class* dari masing-masing kata pada data tweet. Gambar 4 diatas merupakan hasil dari anotasi data.

Data yang telah dilakukan anotasi, kemudian diekstraksi fitur. Pemberian anotasi dan ekstraksi fitur agar memudahkan sistem untuk melakukan klasifikasi.

3.5 Ekstraksi Fitur

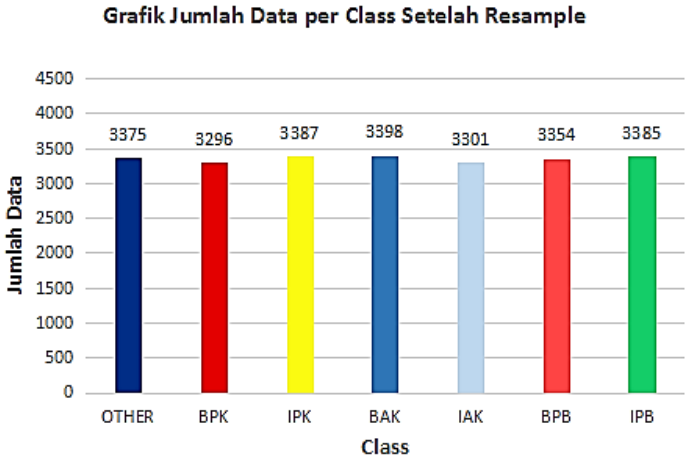
Data yang diektraksi fitur adalah data yang telah dilakukan *preprocessing* dan anotasi data. Ekstraksi fitur dilakukan dengan menggunakan bahasa pemrograman Java, serta menggunakan library InaNLP.jar untuk mendapatkan fitur-fitur NLP. Hasil ekstraksi fitur disimpan ke dalam file format ARFF. Dalam file tersebut ditambahkan *header* berupa daftar atribut dan deklarasi setiap fitur dari masing-masing *token.*

Hasil ekstraksi fitur didapatkan 11748 *token*. Dari hasil tersebut terjadi ketidak-seimbangan dataset (*imbalance dataset*), hal ini disebabkan karena perbedaan jumlah antar kelas sangat besar. Kelas dengan jumlah data yang sangat besar adalah kelas OTHER. Hal ini diketahui dengan bantuan perangkat lunak Weka. Gambar 5 berikut merupakan perbandingan data antar kelas.



Gambar 5. Jumlah Data per Class

Gambar 5 menunjukkan bahwa kelas OTHER memiliki jumlah data yang sangat besar. Untuk mengatasi masalah ini dilakukan *filtering* berupa *resample*. *Resample filtering* dilakukan dengan menggunakan bentuan *library* weka.jar. *Resample filtering* ini akan menyeimbangkan jumlah data per Class, dengan melakukan *OverSampling* dan *UnderSampling* pada data. Gambar 6 berikut menunjukkan jumlah data per Class setelah dilakukan *resample filtering*.



Gambar 6. Jumlah Data per Class Setelah Resample

Gambar 6 menunjukkan keseimbangan data per *class*. Jumlah data per Class sudah tidak terdapat perbedaan yang besar, sehingga hal ini bisa membantu perhitungan klasifikasi menjadi lebih akurat.

3.6 Pembentukan Model

Pembentukan model dilakukan oleh sistem dengan bantuan *library* weka.jar setelah data diseimbangkan. Pembentukan model menggunakan data train yang diambil dari dataset yang yang sudah terbentuk, kemudian mengaplikasikan algoritma *Naive Bayes* ke dalam data train tersebut. Dari proses pembentukan model, sistem menghasilkan keluaran 1 buah file model. File model digunakan sebagai acuan sistem untuk melakukan klasifikasi data uji.

3.7 Klasifikasi Naive Bayes

Proses klasifikasi dilakukan dengan mengambil sebuah data tweet, yang kemudian sistem mencari informasi dari tiap kata yang terdapat di dalam data tweet yang dijadikan sebagai data uji tentang penyakit (BPK dan IPK), akibat (BAK dan IAK), dan penyebab (BPB dan IPB); dan untuk kata yang tidak memiliki informasi diklasifikasikan ke Class OTHER. Tabel 3 berikut merupakan hasil klasifikasi sistem.

Tabel 3. Hasil klasifikasi Sistem

Tweet	Formalisasi	Kata	Prediksi
@rlzapp, Gara gara susu pisang jadi batuk	gara gara susu pisang jadi batuk	gara	BPB
		gara	BPB
		susu	IPB
		pisang	IPB
		jadi	BAK
		batuk	BPK

Tabel 3 menunjukkan hasil prediksi sistem klasifikasi tiap kata yang didapat dari data uji. Data uji berupa tweet mentah yang kemudian diformalisasi tiap kata-nya agar sistem dapat mengklasifikasikan sebagai penyakit (BPK dan IPK), akibat (BAK dan IAK), atau penyebab (BPB dan IPB), dan juga OTHER untuk kata yang tidak memiliki informasi.

3.8 Pengujian

Tabel 4. Perbandingan Hasil Klasifikasi dan Hasil Akurasi

Tweet	Formalisasi	Kata	Aktual	Prediksi	Ket
@rlzapp, Gara gara susu pisang jadi batuk	gara gara susu pisang jadi batuk	gara	BPB	BPB	T
		gara	IPB	BPB	F
		susu	IPB	IPB	T
		pisang	IPB	IPB	T
		jadi	OTHER	BAK	F
		batuk	BPK	BPK	T
@dewicyntia1011, Suara pun mulai serak gara gara batuk :(	suara pun mulai serak gara gara batuk	suara	BAK	BAK	T
		pun	IAK	OTHER	F
		mulai	IAK	IAK	T
		serak	IAK	IAK	T
		gara	OTHER	BPB	F
		gara	OTHER	BPB	F
		batuk	BPK	BPK	T
Jumlah Token			837		
Klasifikasi Benar			620		
Klasifikasi Salah			216		
Akurasi			74,07%		

Pengujian sistem merupakan tahap evaluasi sistem ekstraksi informasi berbasis klasifikasi. Pengujian dilakukan untuk mengetahui akurasi dari model Naive Bayes dan akurasi dari data uji yang diujikan. Pengujian sistem terdapat 2 pengujian, yaitu:

a. Pengujian Model

Pengujian model dilakukan oleh sistem dengan bantuan *library* weka.jar dengan menggunakan metode pengujian 10-*folds Cross-Validation*. Hasil dari pengujian model *Naive Bayes* mendapatkan nilai akurasi 77.27%, presisi 77,8%, dan recall 77,3 %. Berdasarkan hasil pengujian tersebut, dapat dinyatakan bahwa model klasifikasi yang dibentuk dengan algoritma *Naive Bayes* dapat dijadikan acuan untuk klasifikasi karena hasil perhitugan akurasi di atas 70%.

b. Pengujian Klasifikasi

Pengujian klasifikasi dilakukan dengan memasukkan 100 data tweet, untuk kemudian diklasifikasikan oleh sistem untuk mendapatkan informasi tentang penyakit (BPK dan IPK), akibat (BAK dan IAK), atau penyebab (BPB dan IPB). Hasil dari klasifikasi sistem kemudian dibandingan dengan hasil klasifikasi manual dan dihitung akurasinya. Tabel 4 diatasmerupakan hasil dari uji klasifikasi serta perbandingan dengan klasifikasi manual.

Tabel 4 merupakan hasil klasifikasi dan perhitungan akurasi dari 100 data uji yang diujikan. Hasil pengujian didapatkan 837 *token* atau kata yang diklasifikasikan. Hasil klasifikasi menunjukkan 620 *token* diklasifikasikan benar atau *true* (T) dan 216 *token* diklasifikasikan salah atau *false* (F). Perhitungan akurasi data uji menghasilkan nilai akurasi 74,07%.

4. Kesimpulan

Ekstraksi informasi dari twitter dapat dilakukan dengan metode Ekstraksi Informasi berbasis klasifikasi dengan algoritma *Naive Bayes* sebagai algoritma pembelajaran sistem. Pengujian model *Naive Bayes* dengan menggunakan metode 10-*fold cross-validation* mendapatkan hasil akurasi akurasi 77.27%, presisi 77,8%, dan recall 77,3 %. Pengujian klasifikasi dengan menggunakan 100 data uji mendapatkan hasil akurasi 74,07 %.

Untuk pengembangan selanjutnya, perlu adanya perbaikan pada set fitur, karena masing-masing token yang diekstraksi fitur memiliki fitur yang hampir sama dengan token yang lainnya, terutama fitur-fitur NLP. Perlu juga ditambahkannya data penelitian untuk dijadikan data pelatihan, agar dengan meningkatkan data latih diharapkan kinerja klasifikasi sistem lebih baik.

Daftar Notasi

D : Data dengan class yang belum diketahui
h : Hipotesis data D merupakan suatu class yang spesifik
P(h|D) : probabilitas hipotesis h berdasar kondisi D (posterior probability)
P(h) : probabilitas hipotesis h (prior probability)
P(D|h) : probabilitas D berdasar pada hipotesis h
P(D) : probabilitas dari D

Referensi

[1] "6 Masalah Kesehatan Terbesar di Indonesia." [Online]. Available: https://www.guesehat.com/6-masalah-kesehatan-terbesar-di-indonesia. [Accessed: 25-Mar-2018].
[2] C.-E. A. Winslow, "The Untilled Fields of Public Health," *Science (80-.).*, vol. 51, no. 1306, pp. 23–33, 1920.
[3] P. P. Widodo, *16 Penerapan Data Mining Dengan MATLAB*. Bandung: Prestasi Pustaka, 2013.
[4] S. Kumar, F. Morstatter, and H. Liu, "Twitter Data Analytics," *Springer*, p. 89, 2013.
[5] D. Anggareska and A. Purwarianti, "Information extraction of public complaints on Twitter text for bandung government," *Proc. 2014 Int. Conf. Data Softw. Eng. ICODSE 2014*, 2014.
[6] M. L. Khodra, P. Ayu, A. Insanudin, and M. Megally, "Ekstraksi Informasi Transaksi Online pada Twitter," *Cybermatika*, vol. 1, no. July, pp. 1–4, 2013.
[7] L. N. Wulansari, "Ekstraksi Lokasi Dan Produk Dari Data Transaksi Online Pada Twitter," Universitas Muhammadiyah Malang, 2015.
[8] T. R. Hoens and N. V Chawla, "Imbalanced Datasets: From Sampling to Classifiers," *Imbalanced Learn.*, pp. 43–59, 2013.
[9] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.
[10] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. 2011.