Pemodelan Multilabel Tweet Media Sosial Mahasiswa untuk Klasifikasi Keluhan

Faris Musthafa, Joko Lianto Buliali, dan Victor Hariadi
Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: victor@if.its.ac.id

***Abstrak*—Pada umumnya sistem informasi akademik di sebuah perguruan tinggi memiliki fitur umum bagi dosen untuk memantau proses perkembangan akademik anak walinya secara aktif. Namun jika dosen wali ataupun orang tua tidak melakukan pantauan secara aktif maka mahasiswa wali yang memiliki permasalahan akademik berisiko *drop out* dalam proses evaluasi tingkat 1 universitas karena rendahnya pemahaman dosen terhadap mahasiswa walinya. Tujuan dari penelitian ini adalah membuat rancangan model deteksi keluhan dalam data *tweet* mahasiswa. Aspek keluhan bisa dibagi mennjadi empat kategori: keluhan personal, keluhan subjek, keluhan relasi, dan keluhan institusi. Metode multilabel yang digunakan adalah Binary Relevance dengan pilihan *classifier* Naïve Bayes, Simple Logistic, KStar, Decision Table, dan j48. Berdasarkan hasil pengujian ada berbagai *classifier*, Naïve Bayes memiliki performa tertinggi baik dalam aspek akurasi maupun waktu eksekusi. Hasil implementasi sistem deteksi multilabel keluhan menggunakan classifier Naïve Bayes pada delapan puluh data uji yterhadap label keluhan personal, subjek, relasi, dan institusi memiliki akurasi masing-masing bernilai 76.47%, 75%, 80%, dan 80%. Hasil deteksi multilabel keluhan yang ditemukan berpotensi digunakan lebih lanjut pada konteks yang lebih luas.**

***Kata Kunci*—Deteksi Keluhan, Kegagalan Akademik, Pemodelan Prediksi, Multilabel, Media Sosial.**

I. PENDAHULUAN

SALAH satu upaya untuk meningkatkan kualitas Perguruan Tinggi adalah dengan melakukan evaluasi pada mahasiswa yang terdaftar. Evalusi bertujuan untuk menganalisis performa akademik mahasiswa. Salah satu tujuan dari analisis tersebut adalah untuk melakukan deteksi dini terhadap adanya potensi kegagalan akademik mahasiswa yang terkait. Deteksi dini kegagalan akademik bertujuan untuk mengetahui mahasiswa yang berpotensi mengalami permasalahan akademik secara dini.

Pengelompokan faktor-faktor penentu kesuksesan akademik dibagi menjadi empat kelompok: Faktor personal (motivasi, komitmen, dsb.), faktor subjek (struktur matakuliah, kebijakan, buku teks, dsb.), faktor relasi (teman, anggota studi, keluarga, dsb.), dan faktor institusi (lokasi, dosen, staff, fasilitas, dsb/) [1].

Seluruh faktor tersebut tidak bisa digali hanya dari data akademik saja, akan tetapi melibatkan data lain sebagai pelengkap data akademik. Faktor motivasi, komitmen, serta dukungan keluarga dan rekan tidak dapat dilihat didalam data akademik. Oleh karena itu dibutuhkan media sosial sebagai sumber asal data tersebut[2].

Ada tiga permasalahan yang diangkat pada penelitian ini. Pertama adalah bagaimana merancang dan memodelkan klasifikasi multilabel untuk deteksi keluhan dengan memanfaatkan data media sosial *Twitter* sebagai model input dengan menggunakan metode transformasi multilabelisasi Binary Relevance dan *classifier*-nya. Kedua adalah bagaimana melakukan seleksi, penyaringan, dan prapemrosesan data *tweet* sehingga siap digunakan sebagai data latih. Ketiga adalah bagaimana melakukan deteksi multilabel terhadap kumpulan data uji dengan menggunakan model latih yang telah dibuat.

Berdasarkan permasalahan yang ada, kami mengajukan sistem rancangan memodelkan klasifikasi multilabel dengan memanfaatkan data media sosial *Twitter* sebagai model input. Sistem rancangan akan melakukan deteksi multilabel terhadap data uji dengan menggunakan model klasifikasi multilabelisasi yang telah dibuat.

II. DASAR TEORI

A. Teks Mining

Teks mining merupakan proses menggali data teks yang didapatkatn dari sumber data berupa dokumen (*word, pdf,* kutipan, dsb). Teks mining disebut juga dengan teknik mengekstraksi pola dari sebuah dokumen. Tujuan utama dari teks mining adalah untuk mendapatkan informasi yang berguan dari sekumpulan dokumen dengan melakukan pencarian kata-kata yang dapat mewakili isi dari sebuah dokumen yang kemudian dapat dianalisis. Kegunaan dari penggunaan teks mining adalah sebagai kategorisasi teks dan pengelompokan teks[3].

B. Penggalian Data Media Sosial untuk Memahami Siswa

Permasalahan mahasiswa digali pada data posting media sosial *Twitter* dengan mengumpulkan data posting media sosial mahasiswa Universitas Purdue menggunakan bantuan

hashtag #engineeringProblem. Dari data yang terkumpul, peneliti mengambil sampel data dan menurunkan lima label utama yang menjadi topik permasalahan mahasiswa: beban tugas, kurangnya keterlibatan sosial, emosi negatif, kesulitan tidur, dan isu beragaman yang lain. Berdasarkan pengamatan, didapatkan satu posting *Twitter* dapat jatuh kedalam banyak label permasalahan dari lima label utama yang telah didefinisikan[2].

C. Natural Language Processing (NLP)

Natural Language Processing (NLP) adalah alat untuk melakukan ekstraksi dari representasi teks bebas. NLP umumnya menggunakan konsep liquistok (kata benda, kata kerja, kata sifat, dsb) dan struktur gramatikal (kata ganti preposisi, kata ganti benda, dan kata ganti objek)[4].

inaNLP merupakan NLP untuk Bahasa Indonesia yang dapat digunakan untuk formalisasi teks dalam Bahasa Indonesia[5].

D. Klasifikasi Multilabel dan Binary Relevance

Klasifikasi Multilabel adalah klasifikasi dengan tujuan untuk merancang model yang berfungsi untuk melakukan labelisasi pada masing-masing label secara independen dalam kumpulan banyak label[6].

Binary Relevance adalah metode transformasi dengan membagi tiap label secara terpisah dan mengelompokkan berdasarkan metode klasifikasi dasar yang digunakan. Binary Relevance mengasumsikan independensi masing-masing label sehingga mengabaikan keberadaan korelasi antar label[7].

III. METODE PENELITIAN

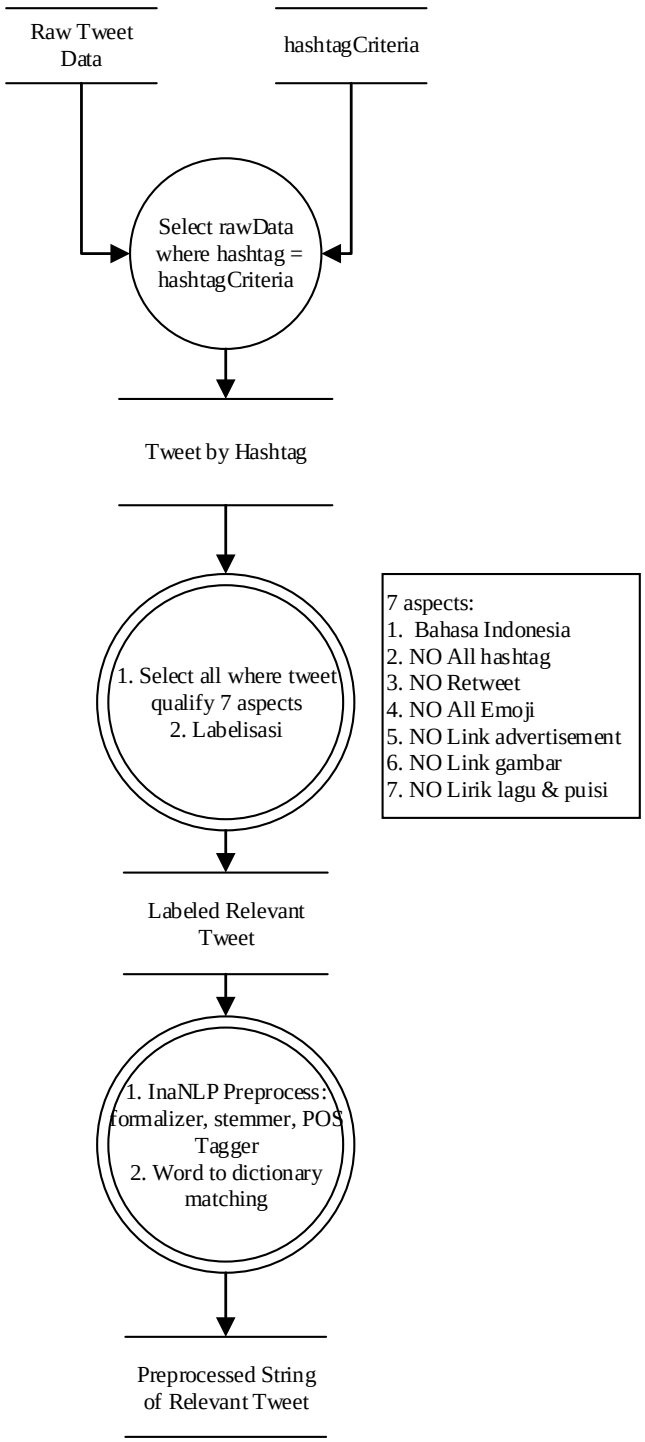
Bab ini akan membahas metode penelitian yang akan dilakukan mulai dari perisapan data yang meliputi tahap prapemrosesan data latih, penentuan *classifier* yang digunakan, pelatihan data latih menggunakan *classifier* pilihan, dan melakukan prediksi menggunakan data uji.

A. Prapemrosesan Data Latih

Prapemrosesan data latih diuraikan pada *data flow diagram* pada Gambar. 1. Prapemrosesan terdiri dari: pengelompokan berdasarkan *hashtag*, eliminasi *tweet* non relevan, labelisasi data latih, dan *formalizer*, *stemmer*, dan POS *Tagger* menggunakan inaNLP.

1) Pengelompokan berdasarkan hashtag

Seluruh data latih *tweet* yang didapat menggunakan API *Twitter* disaring berdasarkan *hashtag* yang relevan dengan cara mengambil *tweet* yang mengandung frasa *hashtag* tertentu. Hal ini bertujuan untuk mempersempit topik bahasan dari jenis *tweet* yang berkategori keluhan yang dialami mahasiswa. Jenis *hashtag* diseleksi berdasarkan kata atau frasa yang memiliki hubungan erat dengan bentuk keluhan yang biasa disampaikan oleh mahasiswa. Terdapat lima puluh lima frasa *hashtag* yang digunakan untuk melakukan penyaringan antara lain: *#penat, #anakteknik, #banyaktugas,* dan *#deritamahasiswa*.



Gambar. 1. *Data flow diagram* prapemrosesan data latih *tweet* mahasiswa

2) Eliminasi tweet non relevan

Pada tahap ini data uji *tweet* yang tidak relevan dieliminasi. Kualifikasi kelulusan relevansi data didasari pada tujuh aspek yaitu: berbahasa Indonesia, tidak berisi *hashtag keseluruhan* dikarenakan tidak adanya konteks kalimat baku, tidak berisi *retweet* dikarenakan *retweet* merupakan bentuk *repost* dari *tweet* pengguna lain yang tidak menggambarkan *tweet* dari user yang diamati, tidak merupakan emoji ataupun emotikon secara keseluruhan dikarenakan tidak adanya konteks kalimat baku, bukan merupakan promosi dan penjualan jasa maupun

produk dikarenakan hal tersebut tidak relevan dengan deteksi keluhan mahasiswa, tidak merupakan gambar dikarenakan penelitian menggunakan teks mining sebagai bahasan sehingga keberadaan gambar tidak relevan, dan bukan merupakan kutipan lirik lagu maupun puisi dikarenakan adanya penggunaan bahasa kiasan atau pergantian makna tanpa ada konteks.

3) Labelisasi Data Latih

Data yang lolos seleksi akan dikasifikasikan menjadi empat label keluhan: keluhan personal, keluhan subjek, keluhan relasi, dan keluhan institusi. Klasifikasi yang dilakukan adalah klasifikasi multilabel, yaitu klasifikasi dengan cara menentukan masing-masing label dengan nilai biner terhadap setiap *data instance*. Setiap *data instance* bisa memiliki satu atau beberapa label keluhan sekaligus.

4) NLP Formalizer

Formalizer merupakan formalisasi atau perbaikan dari kata maupun kalimat yang memiliki *typo* ataupun bahasa tidak baku untuk diganti menjadi kata dengan kemiripan yang tertinggi. Jika kata dengan kemiripan yang cukup tidak ditemukan maka kata tersebut dianggap sebagai kata asing dan tidak mengalami formalisasi.

5) NLP Stemmer

Stemmer berfungsi untuk mengambil kata dasar dari kata berimbuhan yang bertujuan untuk melakukan menghilangkan imbuhan yang tidak relevan pada proses klasifikasi teks.

6) NLP POS Tagger

POS *Tagger* digunakan untuk melakukan ekstraksi kata-kata yang memiliki bobot peranan dan relevansi yang tinggi berdasarkan jenis kategori kata. Kategori kata yang dipilih adalah kata kerja intransitif dan transitif, kata sifat, dan kata benda.

7) Word to word dictionary matching

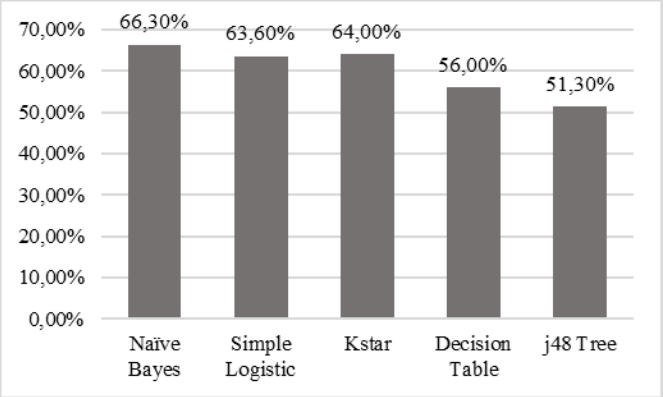
Pada proses POS *Tagger* ditemui masalah berupa lolosnya kata-kata tidak baku yang dianggap sebagai kata benda oleh inaNLP, oleh karena itu dilakukan pencocokan setiap kata benda yang lolos dengan Kamus Besar Bahasa Indonesia dan melakuan eliminasi kata-kata yang tidak ada dalam Kamus Besar Bahasa Indonesia.

B. Penentuan Classifier Latih

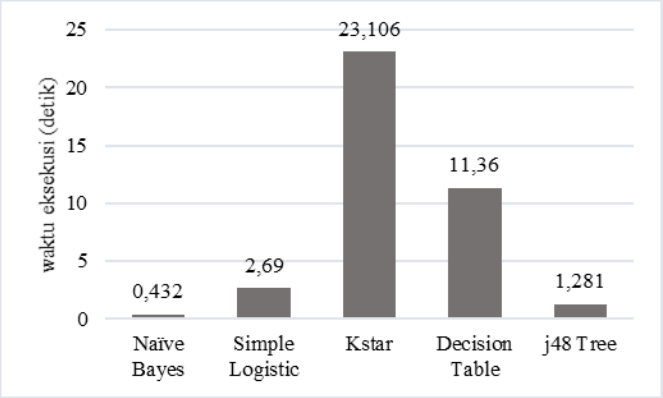
Dalam pertimbangan penetuan jenis *classifier* terbaik dilakukan pengujian *train/test split*, yaitu pengujian performa dari *classifer* menggunakan data latih dan data uji telah di praproses. Pengujian dilakukan dengan cara membagi data latih dan data uji dengan rasio 80% (367 data):20% (92 data) dari keseluruhan 459 data. Pengujian dilakukan pada beberapa jenis *classifier:* Naïve Bayes, Simple Logistic, KStar, Decision Table, dan j48 Tree.

Berdasarkan analisis yang dilakukan, pada Gambar. 2. Naïve Bayes memiliki akurasi sebesar 66.3%, diikuti oleh KStar di urutan kedua dengan akurasi sebesar 64%. Sedangkan berdasarkan Gambar. 3. Berdasarkan aspek waktu eksekusi Naïve Bayes memiliki waktu tercepat 0.432 detik diikuti oleh j48 Tree dengan 1.281 detik.

Berdasarkan dua aspek tersebut *classifier* Naïve Bayes merupakan *classisfier* paling handal baik dalam aspek akurasi dan juga waktu eksekusi dalam penggunaan deteksi multilabel kelujan. Naïve Bayes dipilih sebagai *classifier* pilihan untuk memodelkan klasifikasi multilabel permasalahan mahasiswa.



Gambar. 2. Perbandingan performa akurasi *classifier*



Gambar. 3. Perbandingan waktu eksekusi *classifier*

C. Pelatihan Data Latih

Data *tweet* yang telah dipraproses akan dilatih menggunakan metode transformasi multilabelisasi Binary Relevance dengan *classifier* Naïve Bayes untuk membangun model multilabel keluhan permasalahan mahasiswa.

1) Transformasi Binary Relevance

Binary Relevance adalah metode transformasi data permasalahan multilabel menjadi permasalahan satu label. Data hasil transformasi yang sudah berbentuk satu label baru akan dilatih menggunakan *classifier* Naïve Bayes. Proses Binary Relevance berfungsi untuk mengadaptasi data latih pada *classifier*.

2) Pelatihan Menggunakan Naïve Bayes

Teorema Bayes diuraikan pada persamaan 1 yaitu peluang masuknya karakteristik tertentu dalam kelas *c (posterior)* adalah peluang munculnya kelas *c (prior)* dikalikan dengan peluang kemunculan karakteristik sampel pada kelas *c (likelihood)* dibagi dengan peluang kemunculan karakteristik sampel secara global *(evidence)*.

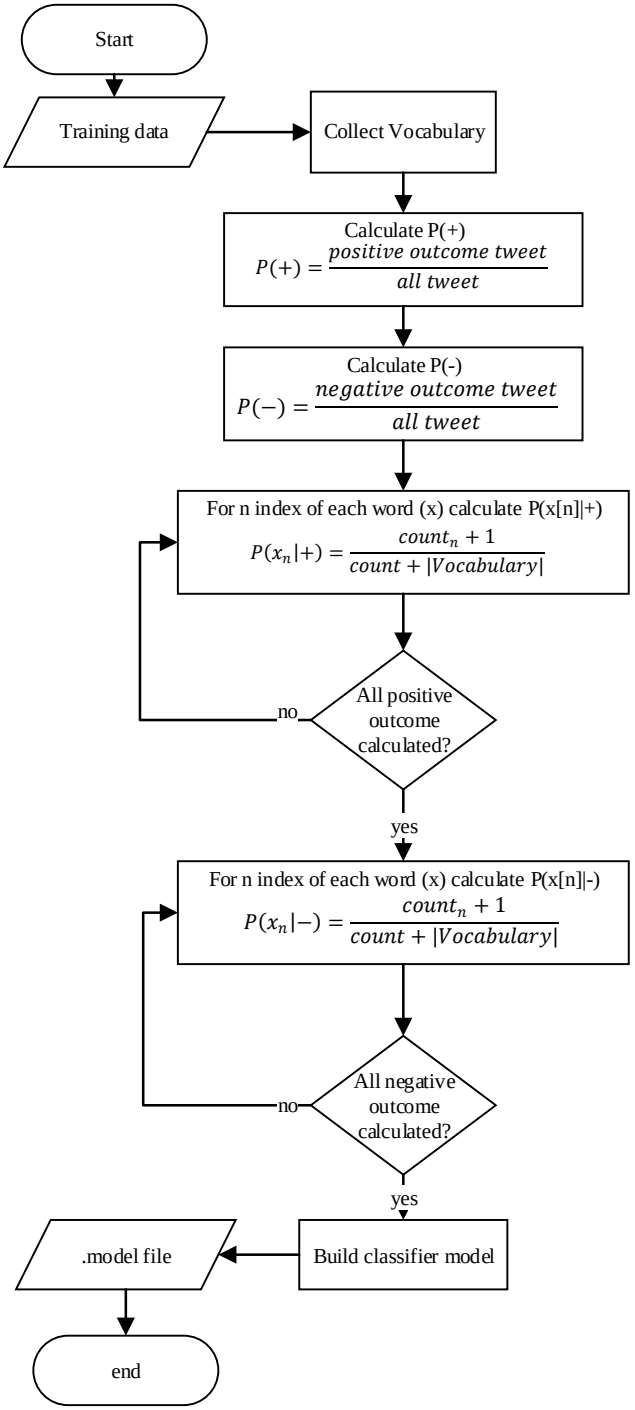
$$P(c|x) = \frac{P(X|C).P(c)}{P(x)} \quad (1)$$

Nilai *evidence* selalu tetap untuk setiap kelas pada satu sampel. Nilai dari *posterior* tersebut nantinya akan

dibandingkan dengan nilai *posterior* kelas lainnya untuk menentukan kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut dari persamaan 1 digambarkan pada persamaan 2.

$$p(c, x_1, \dots, x_n) \quad (2)$$

Hasil penjabaran persamaan 2 bersifat kompleks dan menyulitkan serta proses analisis secara satu persatu yang



Gambar. 4. Proses pelatihan menggunakan *classifier* Naïve Bayes

berakibat pada sulitnya perhitungan yang dilakukan. Hal itu diatasi dengan mengasumsikan independensi *(naive)* dari masing-masing atribut yang terlibat antara satu sama lain. Asumsi tersebut menghasilkan persamaan (3).

$$p(c, x_1, \dots, x_n) = P(c) \prod_{i=1}^{n} P(x_i|c) \quad (3)$$

Gambar. 4 merupakan flowchart proses pembangunan model *classifier* Naive Bayes. Tahap pertama adalah melakukan proses *collect vocabulary* pada data latih yang telah di preproses. *Collect vocabulary* adalah mengumpulkan setiap kata dari seluruh dokumen. Berikutnya adalah menghitung nilai dari *P(+)* yaitu jumlah data *tweet* yang berlabel positif dibagi dengan jumlah data *tweet* keseluruhan. Berikutnya adalah menghitung nilai dari *P(—)* yaitu jumlah data *tweet* yang berlabel negatif dibagi dengan jumlah data *tweet* keselururhan. Tahap berikutnya adalah pembobotan dari masing-masing kata yang ada pada dokumen. Pembobotan label positif dihitung dengan menggunakan persamaan 4 sedangkan pembobotan label negatif dihitung dengan menggunakan persamaan 5. Pembobotan dilakukan pada tiap-tiap kata dengan $count_n$ merupakan jumlah kata yang sedang diberi bobot dalam seluruh data latih, *count(+)* adalah jumlah kata dalam data yang berlabel positif, *count(—)* adalah jumlah kata dalam data yang berlabel negatif, dan *Vocabulary* adalah jumlah kata dari keseluruhan data latih.

$$P(x_n|+) = \frac{count_n+1}{count(+)+|Vocabulary|} \quad (4)$$

$$P(x_n|-) = \frac{count_n+1}{count(-)+|Vocabulary|} \quad (5)$$

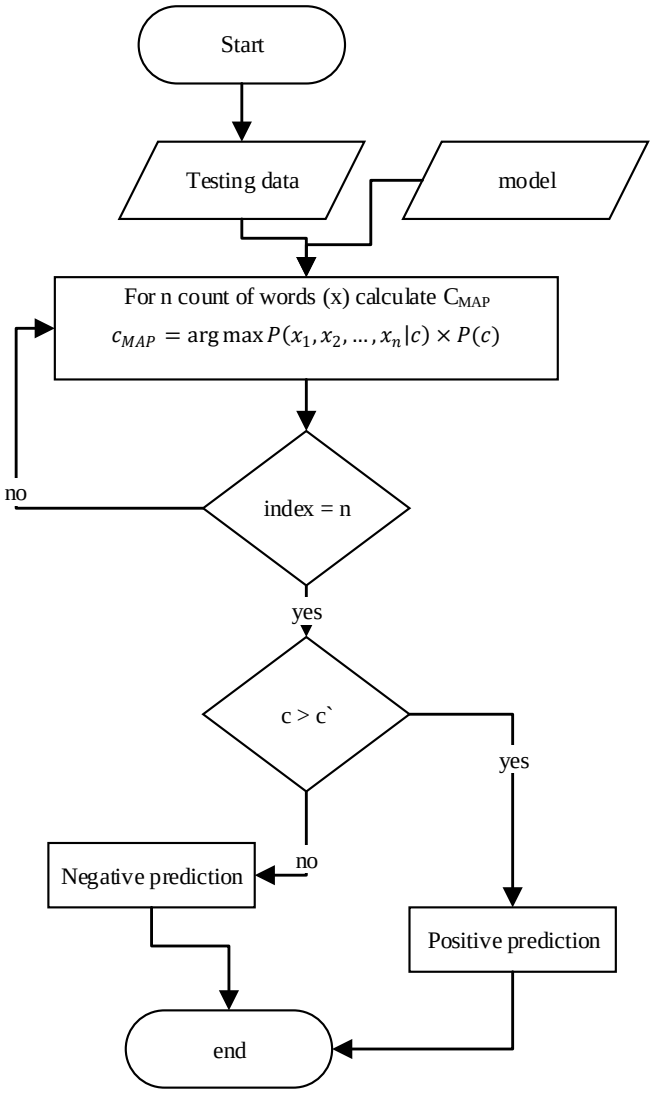
Setelah proses diatas berhasil selesai dilakukan pada satu label selanjutnya dilakukan pengulangan proses pada label-label yang lainnya sehingga mencakup masing-masing label keluhan personal, keluhan subjek, keluhan relasi, dan keluhan institusi. Hasil pembobotan yang sudah dihitung disimpan menjadi rancangan model deteksi multilabel.

D. Prediksi pada Data Uji

Prediksi pada *classifier* Naïve Bayes dihitung dengan menggunakan persamaan 6 yaitu titik dimana fungsi memiliki nilai tertinggi (arg max) dari fungsi probabilitas *outcome* dari masing-masing kata ($P(x_1, x_2, \dots, x_n|c)$) dengan *x* adalah kata yang berkaitan dan *n* adalah jumlah dari seluruh kata dikalikan dengan probabilitas *outcome* ($P(c)$).

$$c_{MAP} = \arg\max P(x_1, x_2, \dots, x_n|c) \times P(c) \quad (6)$$

Gambar. 5 merupakan flowchart proses melakukan deteksi pada data uji dalam suatu label. Penentuan deteksi dilakukan dengan cara membandingkan label postif dengan label negatif menggunakan persamaan 6. Jika berdasarkan persamaan 6 nilai label positif lebih besar dari nilai negatif maka label diklasifikasikan bernilai positif. Sebaliknya jika nilai label positif lebih kecil daripada label negatif maka label diklasifikasikan bernilai negatif. Jika proses deteksi telah selesai dilakukan pengulangan proses deteksi pada label-label lainnya sehingga mencakup masing-masing label keluhan personal, keluhan



Gambar. 6. Proses prediksi menggunakan *classifier* Naïve Bayes

subjek, keluhan relasi, dan keluhan institusi sehingga didapatkan hasil deteksi multilabel.

IV. PENGUJIAN DAN EVALUASI

Uji coba dilakukan dengan menggunakan delapan puluh data *tweet* yang diambil secara acak melaui fitur pencarian *hashtag Twitter*. Data uji kemudian dipraproses dengan menggunakan NLP Formalizer, NLP Stemmer, dan NLP POS Tagger untuk melakukan perbaikan bahasa dan ekstraksi kata-kata relevan.

TABEL 1.
HASIL DARI PENGUJIAN DENGAN MENGGUNAKAN DATA UJI

Akurasi keseluruhan		68.75%
Deteksi benar		55
Deteksi meleset		25
Akurasi	personal	76.47%
	subjek	75.00%
	relasi	80.00%
	institusi	80.00%
	bukan keluhan	23.08%

Berdasarkan statistik pada Tabel 1, uji coba berhasil dilakukan dengan nilai akurasi sebesar 68.75% dengan rincian lima puluh lima deteksi bernilai benar dan dua puluh lima deteksi bernilai salah dari delapan puluh data uji. Akurasi yang didapat dari masing-masing label sebesar 76.47% untuk deteksi keluhan personal, 75% untuk deteksi keluhan subjek, 80% untuk deteksi keluhan relasi, 80% untuk deteksi keluhan institusi, dan 23.08% untuk deteksi non keluhan

V. KESIMPULAN

Berdasarkan pengujian yang telah dilakukan, *classifier* Naïve Bayes menghasilkan deteksi multilabel dengan akurasi teritnggi dan waktu eksekusi terendah dibandingkan dengan metode *classifier* Simple Logistic, KStar, Decision Table, dan j48 Tree.

Pemodelan deteksi multilabel keluhan menggunakan *classifier* Naïve Bayes pada delapan puluh data uji coba terhadap label keluhan personal, subjek, relasi, dan institusi telah berhasil dengan masing-masing akurasi bernilai 76.47%, 75%, 80%, dan 80%.

Model dibuat dengan melakukan prampemrosesan data latih menggunakan NLP *Formalizer*, NLP *Stemmer*, NLP POS *Tagger*, dan *word to word matching* dengan Kamus Besar Bahasa Indonesia. Data yang telah di praproses kemudian dilabelisasi dan dilatih menggunakan transformasi multilabel Binary Relevance dan *classifier* Naïve Bayes. *Word to word matching* dengan Kamus Besar Bahasa Indonesia pada data latih digunakan untuk mengatasi kendala pada NLP POS *Tagger* yang tidak bias membedakan kata ganti benda dengan kata yang tidak baku.

UCAPAN TERIMA KASIH

Penulis F.M. mengucapkan terima kasih kepada semua pihak yang telah ikut serta membantu proses penelitian ini khususnya Departemen Informatika Institut Teknologi Sepuluh Nopember.

DAFTAR PUSTAKA

[1] A. Zhang and C. L. Aasheim, "Academic Success Factors: An IT Student Perspective," J. Inf. Technol. Educ. Res., vol. 10, no. 1, 2011, pp. 309–331.

[2] X. Chen, M. Vorvoreanu, and K. P. C. Madhavan, "Mining Social Media Data for Understanding Students" Learning Experiences," IEEE Trans. Learn. Technol., vol. 7, no. 3, 2014, pp. 246–259.

[3] Cohen KB, Hunter L. "Getting Started in Text Mining". Troyanskaya O, ed. PLoS Computational Biology. 2008.

[4] Daniel Jurafsky and James H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition" (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.A, 2000.

[5] Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif and F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), George Town, 2016, pp. 1-5..

[6] Francisco Herrera, Francisco Charte, Antonio J. Rivera, and Mara J. del Jesus, "Multilabel Classification: Problem Analysis, Metrics and Techniques" (1st ed.). Springer Publishing Company, Incorporated, 2016.

[7] Jesse Read, Peter Reutemann, Bernhard Pfahringer, Geoff Holmes. MEKA: A Multi-label/Multi-target Extension to Weka. Journal of Machine Learning Research. Vol. 17(21). pp 1—5. 2016.