KLASIFIKASI *SPAM E-MAIL* MENGGUNAKAN *METODE TRANSFORMED COMPLEMENT NAÏVE BAYES* (TCNB)

Hanna Florenci Tapikap[1], Bertha S. Djahi[2], Tiwuk Widiastuti[3]
Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana
Email: lauratapika93@gmail.com[1], bertha.djahi@staf.undana.ac.id[2], tiwukwidiastuti@staf.undana.ac.id[3]

INTISARI

Salah satu metode klasifikasi teks yang terkenal adalah metode *Naïve Bayes*. *Naive Bayes* mempunyai komputasi yang efisien serta hasil prediksi yang baik namun performasi *Naïve Bayes* kurang baik saat klasifikasi teks pada *dateset* yang tidak seimbang. Metode *Naïve Bayes ini* kemudian dimodifikasi untuk mengatasi kekurangan tersebut, yang dikenal dengan metode *Transformed Complement Naïve Bayes* (TCNB). Dalam penelitian ini, Metode TCNB digunakan untuk mengklasifikasi *spam e-mail* yang *dataset*-nya tidak seimbang yaitu 481 *dataset* pada *class spam e-mail,* dan 2412 *dataset* pada *class* legitimate *e-mail (*total 2893 *dataset*). Klasifikasi dilakukan dengan *cross validation* dan tanpa *cross validation*. Klasifikasi dengan *cross validation* dilakukan mulai dari *k*=2 sampai dengan *k*=10. Klasifikasi tanpa *cros validation* dilakukan dengan membagi data *training* sebesar 80% dan data *testing* 20%. Hasil menunjukkan klasifikasi menggunakan TCNB dengan *cross validation* mempunyai tingkat akurasi terbaik pada *k*=10 sebesar 93,917% dan klasifikasi tanpa *cross validation* mempunyai tingkat akurasi sebesar 92,760% sehingga dapat dikatakan metode TCNB mampu menangani *dataset* tidak seimbang dengan prediksi akurasi yang baik

Kata kunci*: *Klasifikasi teks, Naïve Bayes, Transformed Complement Naïve Bayes (TCNB), Spam, Legitimate, K-Fold Cross Validation

ABSTRACT

One of the famous text classification methods is the Naïve Bayes Method. Naïve Bayes has efficient computation and good prediction result however the performance of Naïve Bayes is not really good in classifying unbalanced dataset. This Naïve Bayes method is then modified to overcome the weakness, this modified method is then known as Transformed Complement Naïve Bayes (TCNB) method. In this research, TCNB method was used to the spam e-mails whose dataset were unbalanced and were consisted of 481 *dataset* in spam e-mail class, and 2412 dataset in legitimate e-mail class (in total, there are 2893 dataset). The classification was done with and without cross validation. The classification with cross validation was done starting from k=2 until k=10. The classification without cross validation was done by dividing the training data by 80% and testing data by 20%. The result showed that the classification by using TCNB with cross validation had its best accuracy level on k=10 by 93,917% and the classification without cross validation had its best accuracy by 92,760%. Thus it can be concluded that TCNB can handle unbalanced dataset with good prediction accuracy.

***Key Words*: *Text Classification*, *Naïve Bayes*, *Transformed Complement Naïve Bayes* (TCNB), *Spam, Legitimate*, *K-Fold Cross Validation*.**

I. PENDAHULUAN

Spam e-mail adalah *unsolicited automated e-mail* yaitu *e-mail* yang tidak diinginkan yang dikirim secara otomatis (Graham, 2002). Untuk itu perlu adanya *filter e-mail* atau penyaringan *e-mail* sehingga penerima tidak perlu mendapatkan *spam e-mail.* Salah satu cara untuk memfilter *e-mail* adalah dengan teknik klasifikasi dan *naïve bayes* adalah metode yang sering digunakan untuk klasifikasi karena kehandalan dalam klasifikasi teks (Turhan dkk., 2009). *Naïve Bayes* dapat mengklasifikasi beberapa data berupa data teks dengan menerapkan aturan *Bayes*. *Naïve Bayes* mampu memprediksi dan memberikan akurasi yang baik (Turhan dkk., 2009) namun *Naïve Bayes* memiliki kelemahan pada *dataset* yang tidak seimbang. Hal ini yang mempengaruhi hasil performasi *Naïve Bayes* (Sun dkk., 2007).

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

Dataset yang tidak seimbang adalah *dataset* yang memiliki jumlah data *class* tertentu yang jauh lebih dominan dari *class* lainnya (Pozzolo dkk., 2012). Dampak kinerja *Naïve Bayes* saat klasifikasi teks pada *dataset* tidak seimbang adalah pembagian kelas yang salah dengan menempatkan data uji (data *testing*) ke dalam kelas yang dominan. Cara untuk mengatasi masalah *dataset* tidak seimbang yaitu: (i) membuat *dataset* seimbang terlebih dahulu dan (ii) memodifikasi metode *Naïve Bayes* seperti pada metode *Transformed Complement Naïve Bayes* (TCNB).

Metode TCNB dapat mengatasi *dataset* tidak seimbang (Kibriya, 2008). Beberapa tahap dalam metode TCNB, yang tidak ada pada metode *Naïve Bayes* misalnya tahap *Term Frequency* dan *Invers Document Frequency*, bertujuan untuk menentukan bobot pada tiap fitur yang muncul dari setiap dokumen (Rennie dkk., 2003). Klasifikasi menggunkan metode TCNB pernah dilakukan oleh peneliti sebelumnya pada *dataset* teks SMS yang tidak seimbang (Sanu, 2016). Hasil akurasi yang diperoleh sebesar 51,250%. Hal ini dapat disebabkan karena *dataset* yang tidak formal atau kualitas data yang tidak baik, sehingga mempengaruhi hasil akurasi.

Oleh karena itu perlu dilakukan penelitian dengan *dataset* yang berbeda untuk melihat akurasi atau performa dari metode TCNB ini. Pada penelitian ini, penulis melakukan klasifikasi *dataset* tidak seimbang dengan data yang formal yaitu data *e-mail* yang diambil dari corpus Lingspam, menggunakan metode *Transformed Complement Naïve Bayes* (TCNB) untuk melihat tingkat akurasi TCNB.

II. MATERI DAN METODE

2.1 *Naïve Bayes*

Naïve Bayes merupaka metode yang digunakan untuk mengklasifikasikan sekumpulan dokumen. Algoritma ini memanfaatkan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Dalam Anugroho (2010), Pantel dan Lin, serta *Microsoft Research* memperkenalkan metode statistik bayesian ini pada teknologi anti *spam filtering*.

Dalam *naïve Bayes,* kemungkinan dokumen d berada di *class* c dihitung sebagai berikut (Manning dkk., 2009):

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \quad \text{..... (2.1)}$$

dimana:

$P(t_k|c)$ = *conditional probability* dari fitur t_k yang terdapat dalam dokumen dari *class* c.

$P(t_k|c)$ = ukuran berapa banyak kemunculan fitur t_k memberikan kontribusi bahwa c adalah *class* yang benar.

Tujuan utama dalam klasifikasi teks adalah menemukan *best class* untuk sebuah dokumen. *Best class* dalam *naïve Bayes* adalah yang paling mungkin atau *maximum a posteriori* (MAP) *class* c_{map}:

$$c_{map} = \arg max_{c \in C} \hat{P}(c|d) = \arg max_{c \in C} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c) \quad \text{.....(2.2)}$$

Dimana:

$\arg max$ = *argument maximum*

P ditulis $\hat{P}$ karena tidak diketahui nilai sebenarnya dari parameter $P(c)$ dan $P(t_k|c)$.

2.2 *Preprocessing E-mail*

Tahap *pre-processing e-mail* yang digunakan untuk menghilangkan fitur-fitur yang tidak penting dan membersihkan seluruh dokumen *e-mail* dari *markup tag* sehingga saat digunakan akan meningkatkan akurasi dari *classifier*.

1. Tokenisasi

   Tokenisasi adalah mengekstrak isi dokumen *e-mail* ke dalam bentuk kata/fitur. Pada tahap ini juga dilakukan *case folding*, yaitu penyeragaman bentuk huruf "a" sampai huruf "z" dan juga menghilangkan karakter-karakter tertentu seperti tanda baca dan angka.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

2. Eliminasi *stopwords*
   Stopwords adalah kata-kata yang memiliki frekuensi kemunculan yang tinggi dalam dokumen namun tidak memiliki nilai informasi yang tinggi. Contoh *stopword* misalnya kata "*the", "and*", "*a*", "*of*", "*in*", "*is", "this*", "*that*".
3. Lemmatisasi fitur
   Proses lemmatisasi (*lemmatization*) akan mereduksi kata ke dalam bentuk kata dasar atau sering disebut lemma dimana kata dasar (lemma) tersebut memiliki arti dan ditemukan dalam kamus. Misalnya, kata "*walked*", "*walks*" dan "*walking*" jika dilakukan proses lemmatisasi maka lemma-nya adalah "*walk*" (kata "*walk*" ada di dalam kamus Bahasa Inggris).

2.3 *Dataset* Tidak Seimbang

Masalah utama dalam klasifikasi teks adalah *dataset* yang tidak seimbang, dimana jumlah *class* yang satu lebih dominan dari *class* yang lain. Efek negatif daripada ketidakseimbangan ini adalah *classifier* tidak dapat memprediksi data dengan benar karena *classifier* akan menghasilkan hasil prediksi untuk *majority class* sangat baik tetapi buruk saat memprediksi *minority class*.

Pada *dataset Corpus Lingspam* yang terdiri dari dua *class* yaitu *spam* dan *legitimate* dimana jumlah dokumen *spam* adalah 481 dan jumlah dokumen *legitimate* adalah 2412. Hasilnya menunjukan *class spam* memiliki nilai *f1-measure* 17,45% dan *class legitimate* memiliki nilai *f1-measure* 91,73%.

2.4 Solusi Masalah Dataset Tidak Seimbang

Untuk mengatasi masalah *dataset* tidak seimbang, ada dua cara pendekatan yang dilakukan; (i) mengimbangkan *dataset* terlebih dahulu dengan teknik *sampling* pada *traning dataset*. Teknik ini membuat *class majority* dimodifikasi sedekimian sama dengan *class minority*, dan (ii) menambahkan beberapa tahap atau langkah dalam klasifikasi. Langkah awal pada *training dataset* dibiarkan tidak seimbang tetapi algoritmanya yang dimodifikasi dari *metode Naïve Bayes*menjadi metode *Transformed Complement Naïve Bayes* (TCNB). Ada beberapa langkah dalam metode TCNB yaitu *Term Frequency, Invers Document Frequency, Length Normalized, dan Complement Naïve Bayes.*

2.5 Transfromed Complement Naïve Bayes (TCNB)

Naïve Bayes memiliki kesalahan yang sistemik yaitu penyimpangan data, dimana hasil algoritma *naïve Bayes* pada saat menangani *dataset* yang tidak seimbang mengakibatkan pembagian *class* yang salah yaitu menempatkan data baru (data *testing*) ke dalam *class* yang dominan.

2.5.1 Term Frequency (TF)

Term Frequency (TF) yaitu untuk menentukan bobot kata pada suatu dokumen berdasarkan jumlah kemunculan kata dalam dokumen tersebut. Semakin besar jumlah kemunculan suatu kata (TF) dalam dokumen, semakin besar pula bobotnya dalam dokumen. Persamaan TF sebagai berikut:

$$d_{ij} = log(d_{ij} + 1) \quad \text{.....(2.3)}$$

Dimana: d_{ij} = jumlah kemunculan kata i dalam dokumen j.

2.5.2 Invers Document Frequency (IDF)

Invers document frequency (IDF), yang digunakan untuk memperhitungkan kata-kata yang sering hadir dalam semua dokumen. Secara umum dapat digunakan persamaan sebagai berikut:

$$d_{ij} = \log(f + 1) \times \log\left(\frac{D}{d_f}\right) \quad \text{.....(2.4)}$$

Dimana: D = jumlah seluruh dokumen.

d_f = jumlah dokumen yang mengandung kata i.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

f = dij nilai TF *transform.*

2.5.3 Length Normalized (LN)

Penggunaan frekuensi *term* dalam dokumen sebagai bobot *term* dalam representasi dokumen tidaklah memadai. Hal ini karena bias dapat muncul dari faktor lain, misalnya banyaknya dokumen yang memuat *term* tersebut, atau faktor panjang dokumen dimana *term* tersebut muncul. Untuk itu normalisasi frekuensi *term* terhadap panjang dokumen sangat diperlukan. Persamaannya sebagai berikut:

$$d_{ij} = \frac{d_{ij}}{\sqrt{\Sigma_k (d_{kj})^2}} \quad \text{.....(2.5)}$$

dimana:

dij = nilai dari IDF

$\Sigma_k (d_{kj})$ = jumlah seluruh kata yang ada di dokumen *j*.

2.5.4 Complement Naïve Bayes

Untuk menghadapi penyimpangan *training dataset*, maka akan digunakan sebuah "kelas pelengkap" dalam versi *naïve Bayes* yang disebut metode *complement naïve Bayes* (CNB). Perkiraan parameter *complement naïve Bayes* (CNB) menggunakan data dari seluruh *class* $\tilde{c}$. Persamaan CNB sebagai berikut:

$$\widehat{\theta}_{\tilde{c}i} = \frac{N_{\tilde{c}i} + \alpha i}{N_{\tilde{c}} + \alpha} \quad \text{.....(2.6)}$$

dimana:

$N_{\tilde{c}i}$ = jumlah kata i dalam dokumen pada *class* $\tilde{c}$

$N_{\tilde{c}}$ = jumlah seluruh kata dalam semua dokumen pada *class* $\tilde{c}$.

$\alpha_i = 1$

α = jumlah kata yang tidak berulang atau unik (*vocabulary*).

Untuk menghitung bobot setiap fitur dalam sebuah dokumen pada masing-masing *class* digunakan persamaan sebagai berikut:

$$\widehat{w}_{\tilde{c}i} = \log \widehat{\theta}_{\tilde{c}i} \quad \text{.....(2.7)}$$

sehingga persamaan untuk memberikan label pada dokumen uji adalah sebagai berikut:

$$l(t) = argmin_c \sum i \, t_i w_{ci}$$

dimana: .....(2.8)

$\sum i$ adalah frekuensi kata i dan $(t_i w_{ci})$ adalah nilai w_{ci} dari kata i

2.6 K-Fold Cross Validation

K-Fold Cross Validation merupakan salah satu metode dalam menentukkan data *training* dan data *testing* dari keseluruhan data. Data yang digunakan di secara acak ke dalam *k subset* yaitu *D1,D2,...*D_k dengan ukuran yang sama. Proses *training* dan *testing* dilakukan sebanyak *k*-kali secara berulang-ulang. Pada iterasi ke-i, partisi D_i disajikan sebagai data *training*. Iterasi kedua, subset $D_1, D_2, \ldots . D_k$ akan dites pada D_2, dan selanjutnya hingga D_k. (Han*, et al,* 2012:364).

III. HASIL DAN PEMBAHASAN

3.1 Klasifikasi menggunakan TCNB

Tahap klasifikasi menggunakan *Transfomed Complement Naïve Bayes* (TCNB) dimulai perhitungan sebagai berikut:

1. Tahap perhitungan *Term Frequency* (TF) menggunakan Persamaan (2.3) dengan mencari fitur dalam dokumen, menghitung jumlah kemunculan setiap fitur.
2. Tahap perhitungan *Invers Document Frequency* (IDF) menggunakan Persamaan (2.4) dengan mencari nilai *tf* dari masing-masing fitur, hitung total seluruh dokumen dan jumlah dokumen yang memiliki fitur yang sedang diperiksa.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

3. Hitung *Length Normalization* (LN) dengan mencari nilai *idf* dari setiap fitur jumlah seluruh nilai *idf* dari masing-masing fitur yang ada dalam dokumen dan menggunakan Persamaan (2.5).
4. Hitung nilai *Complement Naïve Bayes* dilakukan pencarian jumlah *vocabulary*, cari nilai *length normalization* pada setiap fitur di kelas, hitung nilai *complement* menggunakan Persamaan (2.6) dan *log complement* pada *class* menggunakan Persamaan (2.7). Hasil hitung nilai *complement* dalam menentukan *class Spam, class Ham* dan juga merupakan hasil klasifikasi TCNB

3.2 Pengujian

3.2.1. Single Testing

Dalam *Single Testing*, pertama dilakukan adalah *upload file*. *File* atau *e-mail testing* dimaksudkan adalah *file* baru yang belum diklasifikasikan.

3.2.2 *K-Fold Cross Validation*

Data yang digunakan dalam pengujian *K Fold Cross Validation* sebanyak 2893 data. Langkah awal dalam pengujian ini adalah dengan membagi data menjadi *k-fold* bagian yang jumlahnya sama atau seimbang. Hasil pembagian *k fold* digunakan setiap *fold* untuk data *testing* dan sisanya untuk data *training*. Pengujian *K-Fold* parameter menggunakan *K-Fold*= 2, *K-Fold*= 3, *K-Fold*= 4, *K-Fold*= 5, *K-Fold*= 6, *K-Fold*= 7, *K-Fold*= 8, *K-Fold*= 9, dan *K-Fold*= 10.

3.2.3 Pengujian *tanpa K-Fold Cross Validation*

Pengujian *without K-Fold* (Tanpa K-Fold) menggunakan 80% dan 20%. Dalam pengujian ini, dimulai membagi data 80% adalah data latih dan 20% adalah data uji dengan 2893 data. Rincian pembagian data 80% dan 20% sebagai berikut;

$$\text{Pembagi data 80\% (data latih)} = \frac{2893}{80\%} \text{x } 100\% = 2314 \text{ data}$$

$$\text{Pembagi data 20\% (data uji)} = \frac{2893}{20\%} \text{x } 100\% = 578 \text{ data}$$

Hasil pembagian data tersebut, pengujian melakukan sebanyak 10 kali uji. Hasil pengujian 80% dan 20%, akan dihitung nilai akurasi dengan rumus;

$$\text{Akurasi} = \frac{Jumlah\ data\ benar\ yaitu\ TP+TN}{total\ data} x\ 100\%$$

IV. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Pada penelitian ini , pengujian menggunakan *K-fold cross validation* dan Tanpa *K-fold* dilakukan sebanyak 10 kali untuk memprediksi keakuratan hasil klasifikasi TCNB. *K-fold cross validation* terbaik saat *k-fold* =10 dengan akurasi 93.917% dimana perbandingan data latih dan data uji adalah 9:1, sedangkan tanpa *k-fold* dengan 80% data latih dan 20% data uji dari total 2893 dataset diperoleh akurasi 92.760%. Metode *Transformed Complement Naïve Bayes* (TCNB) dapat mengklasifikasi *spam e-mail* dengan nilai akurasi sebesar 93.917%, dan dapat menangani *dataset* yang tidak seimbang dengan total *Spam e-mail* sebanyak 481 total *Ham* (Legitimate) *e-mail* sebanyak 2412.

4.2 Saran

Diperlukan adanya penelitian lanjutan menggunakan algoritma TCNB untuk klasifikasi *spam e-mail* dengan dataset yang lebih bervariasi untuk melihat tingkat konsistensi tingkat akurasi algoritma TCNB.

DAFTAR PUSTAKA

[1] Aamodt, A., dan Plaza, E., 1994, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, Vol. 7, 39-59.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)

[2] Jackson et al,.1989, *Similarity Coefficient: Measures of co-occurrence and association or simply measures of occurrence,* University of Toronto,Canada.
[3] Graham P., 2002. A Plan for Spam.
[4] Pozzolo, A., Caelen, O. and Bontempi, G., 2012, *Comparison of balancing techniques for unbalanced datasets*.
[5] Sun,Y.,Mohamed, K, S., Wong, A. K., & Wang, Y., *Cost-sensitive Boosting fof Classification of Imbalanced Data*. Pattern Recognition Society, 3358-3378.
[6] Kibriya,A., Frank, E., Pfhringer, B. and Holmes, G.,2008, Multinomil naïve Bayes for text categorization revisited.
[7] Rennie, J., Shih, L., Teevan, J.and Karger, D., 2003, *Tackling the Poor Assumptions of Naïve Bayes Text Classifier.*
[8] Sanu, Anindhyan., 2016, Studi Perbandingan Performansi *Multinominal naïve Bayes* dan *Transformed Complement Naïve Bayes* saat klasifikasi teks pada *Dataset* yang tidak seimbang.
[9] Mahinovs, A. and Tiwari, A., 2007, *Text classification method review*.
[10] Saad, Omar., Darwish, Asharf., and Faraj, Ramadan., 2012. A survey of Machine Learning Techniques for Spam Filtering, *International Journal of Computer Science and Network Security*.
[11] Anugroho, Prasetyo.,2010. Klasifikasi *e-mail spam* dengan metode naïve bayes *classifier* menggunakan *java programming*.
[12] Manning, C., Raghavan, P. and Schutze, H., 2009. *An introduction to information retrieval*.
[13] Full. 1994. Neural Network in Computer Science. Singapura: McGrawHill.
[14] Han, J., and Kamber M. 2006. Data Mining:*Concept and Techniques.* New York:Morgan Kaufmann Publisher.
[15] Sheu, jyh-jian.2008. An Effecient Two-Phase Spam Filtering Method Based On E-mails Categorization.

ISSN:2337-7631 (printed)
ISSN: 2654-4091 (Online)