

## PERBANDINGAN ALGORITMA NAIVE BAYES, ID3, DAN TAN PADA KLASIFIKASI SMS SPAM

<sup>1</sup> Arie Satia Dharma, <sup>2</sup> Oktavi Yanty Silitonga, <sup>3</sup> H.Justin Manurung

<sup>1,2,3</sup> Program Studi Sarjana Teknik Informatika,

<sup>1,2,3</sup>Institut Teknologi Del

email: ariesatia@del.ac.id

**Abstrak.** Short Message Service atau SMS merupakan layanan yang memungkinkan untuk mengirim dan menerima pesan dalam bentuk teks pada perangkat telepon genggam. SMS spam adalah pesan yang tidak diinginkan, yang kita tidak ingin pesan tersebut berada di dalam kotak pesan kita. Salah satu cara untuk mengatasi masalah SMS spam tersebut adalah dengan melakukan klasifikasi terhadap teks SMS untuk menentukan spam menggunakan teknik machine learning. Penelitian ini melakukan eksperimen untuk membandingkan tingkat akurasi penerapan metode supervised learning menggunakan algoritma Naive Bayes (NB) Classifier, Iterative Dychotomizer Version 3 (ID3) dan Tree Augmented Naïve (TAN) Bayes Classifier dalam mengklasifikasikan data SMS. Supervised learning merupakan metode untuk mengklasifikasikan data berdasarkan label yang sudah ada. NB adalah klasifikasi data berdasarkan nilai peluang. ID3 adalah algoritma yang melakukan prosedur pencarian secara menyeluruh kepada semua kemungkinan yang akan terjadi. TAN adalah pengembangan dari NB dimana antar node atribut dapat saling memiliki ketergantungan. Berdasarkan hasil eksperimen yang dilakukan, dengan menggunakan pembagian data 70% data training dan 30% data testing diperoleh bahwa algoritma NB menghasilkan akurasi yang paling tinggi sebesar 96.21% dibandingkan dengan kedua algoritma yang lain. Sementara dengan menggunakan pembagian data 80% data training dan 20% data testing diperoleh bahwa algoritma ID3 menghasilkan akurasi yang paling tinggi sebesar 96.47% dibandingkan dengan kedua algoritma yang lain.

**Kata Kunci:** SMS Spam, Klasifikasi SMS Spam, Machine Learning, Naive Bayes, ID3, Tree Augmented Naive Bayes.

**Abstract.** Short Message Service or SMS is a service that allows sending and receiving messages in text at a mobile device. SMS spam is an unwanted message, which we don't want it to be saved in message box. One of the things that can be done for solving the SMS spam's problem is to do SMS classification to determine spam using machine learning techniques. This researching conducted experiments to compare the accuracy by applying supervised learning methods using the Naive Bayes (NB) Classifier algorithm, Iterative Dychotomizer Version 3 (ID3) algorithm and the Tree Augmented Naïve (TAN) Bayes Classifier algorithm. Supervised learning is a method for classify data based on existing labels. NB is a combined data based on probability. ID3 is an algorithm that performs procedures to achieve what will happen. TAN is the development of NB where each node attributes can be dependent each other. Based on the results of experiments carried out, by using splitting data 70% training data and 30% testing data, obtained that the NB algorithm produces the highest accuracy of 96.21% compared to the two algorithms. While using splitting data 80% training data and 20% testing data obtained that ID3 algorithm produces the highest accuracy of 96.47% compared to the other two algorithms.

**Keywords:** SMS Spam, SMS Spam Classifier, Machine Learning, Naïve Bayes, ID3, Tree Augmented Naïve Bayes.

## PENDAHULUAN

Short Message Service atau SMS merupakan sebuah media komunikasi berbentuk teks yang memungkinkan pengguna ponsel untuk saling berbagi teks (Arifin & Bijaksana, 2016). Meningkatnya komunikasi melalui SMS telah banyak dimanfaatkan oleh pihak-pihak tertentu untuk mengirimkan pesan-pesan yang tidak bermanfaat atau SMS spam (Dewi & Supriyanto, 2013). SMS spam adalah pesan yang tidak diinginkan, yang kita tidak ingin pesan tersebut berada di dalam kotak pesan kita (Sethi & Bhootna, 2014).

Salah satu hal yang dapat dilakukan untuk mengatasi masalah SMS spam adalah dengan melakukan klasifikasi teks SMS menggunakan teknik *machine learning*. Machine learning merupakan salah satu metodologi cerdas yang dapat memberikan hasil yang menjanjikan dalam domain (ruang lingkup) klasifikasi dan prediksi (Bunker & Thabtah, 2017). Teknik klasifikasi merupakan salah satu metode machine learning yang dilakukan dengan cara supervised learning yaitu dengan mengklasifikasikan data berdasarkan arahan yang sudah ada (Liu, 2007). Supervised learning memiliki karakter menyelesaikan masalah berupa klasifikasi.

Metode yang dibahas pada penelitian ini adalah menerapkan metode supervised learning dengan menggunakan algoritma Naïve Bayes (NB) Classifier, Iterative Dichotomizer Version 3 (ID3) dan Tree Augmented Naïve (TAN) Bayes Classifier. Peneliti memilih algoritma NB dikarenakan NB merupakan salah satu algoritma yang efektif diterapkan untuk melakukan klasifikasi dengan jumlah data yang besar dan menghasilkan nilai akurasi yang baik (Dewi & Supriyanto, 2013), dan untuk memilih algoritma ID3 karena algoritma ini memiliki tingkat akurasi yang baik dalam menyelesaikan solusi klasifikasi dengan memanfaatkan teknik klasifikasi data mining (Aradea, Ariyan, & Yuliana, 2011). Sementara untuk memilih algoritma TAN karena TAN merupakan pengembangan dari algoritma NB serta metode tersebut merupakan algoritma yang efektif dan memiliki akurasi yang tinggi dalam mengklasifikasikan data jika data yang dipakai memiliki hubungan antaratribut (Padmanaban, 2014).

Berdasarkan penelitian yang telah dilakukan, ketiga algoritma yang akan diterapkan dalam Tugas Akhir ini masing-masing memiliki akurasi yang tinggi. Akan tetapi, masing-masing algoritma melakukan klasifikasi data dengan cara yang berbeda-beda. Oleh sebab itu, peneliti bermaksud untuk membandingkan masing-masing algoritma dengan menerapkan teknik machine learning.

## METODE PENELITIAN

### Pengumpulan Dataset

Dataset yang dikumpulkan diambil dari *University of California Irvine (UCI) Repository* dan *Dublin Institute*

*of Technology Repository*. Total data yang diambil dari UCI *Repository* adalah 5572 data SMS yang terdiri dari 747 SMS spam dan 4825 SMS non spam. Total data yang diambil dari *Dublin Institute of Technology Repository* adalah 1353 SMS spam. Sehingga total data yang dikumpulkan adalah 6925 dengan 2100 SMS spam dan 4825 SMS non spam. Setelah semua data sudah terkumpul, kemudian isi pesan yang berulang ataupun *redundant* dihapus, karena tidak perlu untuk menggunakan data yang berulang. Dan apabila ditemukan dalam *dataset* data yang hanya mengandung tanda baca, maka data tersebut juga akan dihapuskan. Sehingga semua data menjadi unik. Setelah menghapus data yang berulang, total data yang dikumpulkan adalah 6520 dengan 2005 spam dan 4515 non spam. Berdasarkan pembelajaran yang dilakukan penulis tentang keseimbangan data, jumlah data kelas *spam* dan *non spam* tidak berbeda jauh sehingga *dataset* termasuk data yang *balance* (Batista, Prati, & Monard, 2004). Dengan komposisi data seperti ini, *dataset* tersebut memiliki rasio perbandingan data 1:2. Oleh karena itu, *dataset* tersebut digolongkan *balance* ataupun seimbang. Selanjutnya data tersebut akan disimpan dalam bentuk file yang berekstensi *.csv*.

### Preprocessing Dataset

Pada tahapan *preprocessing* terdapat beberapa proses yang dilakukan yaitu *case folding*, *non-alphabet removal*, *stopword removal*, dan *stemming* (Arifin & Bijaksana, 2016; Srividhya & Anitha, 2010).

#### 1. Case folding

*Case folding* merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter* (pembatas).

#### 2. Non-alphabet Removal

*Non-alphabet removal* merupakan tahapan yang menghilangkan karakter yang bukan merupakan huruf seperti tanda baca dan angka pada teks sms karena tidak memberikan nilai informasi.

#### 3. Stopword Removal

*Stopword* merupakan kata penghubung dan kata ganti. Beberapa contoh *stopword* pada konten bahasa Inggris diantaranya kata "i", "you", "she", "he", "to", "be" dan sebagainya. *Stopword removal* dilakukan untuk membuang semua *stopword* yang tidak memberikan nilai yang tidak akan mempengaruhi proses *text mining*.

#### 4. Stemming

*Stemming* merupakan proses mengembalikan setiap kata kembali ke bentuk dasarnya dengan menghilangkan semua akhiran yang terdapat pada kata tersebut seperti -ing, -ed, dan -er. Jika dilihat dalam bahasa Inggris kata-kata yang mengandung akhiran adalah seperti, finished,

finishing, finisher. Ketiga kata tersebut akan diubah menjadi 1 kata yang sama yaitu finish.

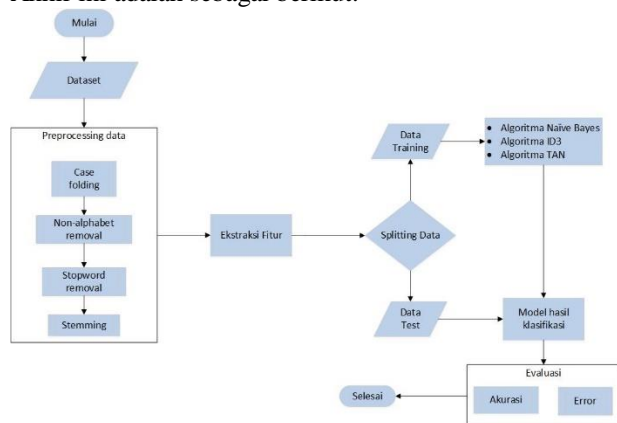
### Feature Extraction

Ekstraksi fitur adalah proses mengubah suatu bentuk data menjadi bentuk numeric. Dalam penelitian ini, feature extraction yang dipakai menggunakan pendekatan bag-of-words model. Pemilihan model ini dikarenakan model ini cocok digunakan untuk merepresentasikan data berupa teks menjadi data numerik (Zhang, Jin, & Zhou, 2010) [19].

Tahap feature extraction dilakukan sebelum melakukan pembagian dataset agar matrix antara data training dan data testing memiliki ukuran yang sama sehingga dapat diproses menggunakan algoritma machine learning.

### Rancangan Eksperimen

Rancangan desain eksperimen pengerjaan Tugas Akhir ini adalah sebagai berikut.



Gambar 1. Rancangan Eksperimen

Keterangan skema rancangan eksperimen:

1. Sistem menerima inputan *file dataset* yang berekstensi .csv.
2. Sebelum digunakan, semua *dataset* akan melakukan tahap *preprocessing dataset* menggunakan fungsi *case folding*, *non-alphanumeric*, *stopword removal* dan *stemming*.
3. Setelah *dataset* di-*preprocessing*, *dataset* tersebut akan diolah menggunakan ekstraksi fitur.
4. Setelah selesai, akan diperoleh *dataset* hasil *preprocessing* yang akan dibagi menjadi data *training* dan data *testing* dengan menggunakan metode *hold-out*. Data yang menjadi data *training* akan berguna untuk melatih sistem untuk dapat mengklasifikasikan data SMS. Sedangkan data *testing* berguna untuk menguji keakuratan dari sistem yang telah dilatih sebelumnya. Pembagian data yang dilakukan adalah sebagai berikut:
  - a. Pembagian data dengan 70% data *training* dan 30% data *testing*
  - b. Pembagian data dengan 80% data *training* dan 20% data *testing*

Pembagian data ini bertujuan untuk seberapa besar pengaruh pembagian data terhadap akurasi yang diperoleh.

5. Lalu, data *training* diolah menggunakan algoritma Naïve Bayes Classifier, ID3, dan Tree Augmented Naïve Bayes Classifier. Sehingga akan diperoleh masing-masing model dari algoritma yang digunakan.
6. Setelah memperoleh model dari ketiga algoritma tersebut, dilakukan tahap evaluasi dengan menghitung akurasi dan *error* dengan menggunakan *confusion matrix*.

### Prototype Pengujian

Pada tahapan ini, peneliti membangun 2 interface untuk melakukan pengujian terhadap masing-masing algoritma yaitu membangun interface dengan menggunakan bahasa pemrograman Python dan Tkinter serta membangun interface dengan menggunakan bahasa pemrograman R yang terhubung ke Java.

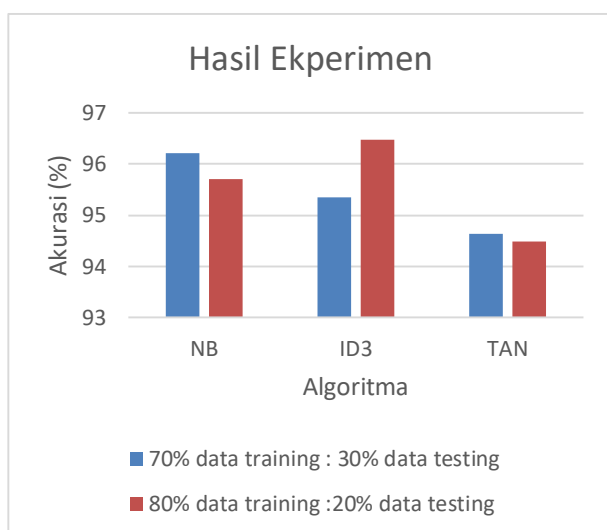
## HASIL DAN PEMBAHASAN

Hasil eksperimen yang dilakukan dengan masing-masing algoritma dapat dilihat pada Tabel 1.

Tabel 1. Hasil Eksperimen

Pembagian Data (%)		Akurasi(%)		
Data Training	Data Test	NB	ID3	TAN
70	30	96.21	95.34	94.63
80	20	95.7	96.47	94.48

Diagram perbandingan akurasi menggunakan algoritma Naïve Bayes (NB) Classifier, Iterative Dychotomizer Version 3 (ID3), dan Tree Augmented Naïve (TAN) Bayes Classifier dengan pembagian data 70% data *training* : 30% data *testing* dan 80% data *training* : 20% data *testing* dapat dilihat pada Gambar 2.



Gambar 2 Diagram Perbandingan Akurasi menggunakan algoritma NB, ID3, dan TAN

Dari hasil eksperimen ini, dapat dilihat bahwa pembagian data berpengaruh terhadap akurasi yang diperoleh dari setiap model. Namun pengaruh pembagian data kecil karena perbedaan akurasi yang diperoleh pada setiap algoritma kecil. Untuk algoritma Naïve Bayes perbedaan akurasi sebesar 0.51%, untuk algoritma ID3 perbedaan akurasi sebesar 1.13%, dan untuk algoritma TAN perbedaan akurasi sebesar 0.15%. Hasil eksperimen menunjukkan bahwa dengan menggunakan pembagian data 70% data *training* dan 20% data *testing* dapat disimpulkan akurasi yang paling tinggi sebesar 96.21% diperoleh dengan menggunakan algoritma Naïve Bayes Classifier. Sementara dengan pembagian data 80% data *training* dan 20% data *testing* dapat disimpulkan akurasi yang paling tinggi sebesar 96.47% diperoleh dengan menggunakan algoritma Iterative Dychotomizer Version 3. Algoritma Naïve bayes dan Tree Augmented Naïve Bayes mengalami penurunan akurasi pada pembagian data 80% data *training* dan 20% data *testing*. Namun Algoritma Iterative Dychotomizer Version 3 mengalami kenaikan pada pembagian data tersebut. Akurasi tertinggi yang didapat pada penelitian ini adalah menggunakan algoritma Iterative Dychotomizer Version 3. Walaupun pada pembagian 70% data *training* dan 30% data *testing*, algoritma ini lebih kecil akurasinya dibandingkan dengan algoritma Naïve Bayes, yang memiliki kompleksitas paling kecil.

Berdasarkan analisis dan eksperimen yang telah dilakukan pada seluruh pengerjaan Tugas Akhir ini memperoleh nilai akurasi yang diatas 90% untuk semua algoritma. Namun, pada tahap *preprocessing* data masih terdapat kata-kata yang tidak ada dalam kamus bahasa Inggris, dan juga terdapat nama *merk* ataupun produk tertentu yang sebenarnya kata-kata tersebut harusnya dihilangkan ataupun dapat diperbaiki. Kata-kata tersebut dapat dilihat pada Lampiran N. Dengan menghapus dan memperbaiki kata-kata tersebut, komputasi yang dilakukan pada algoritma akan lebih kecil dan akan lebih efektif

karena komputasi yang dilakukan hanya untuk kata-kata yang memiliki arti.

## KESIMPULAN

Berdasarkan implementasi yang dilakukan maka dapat ditarik kesimpulan berikut:

1. Penelitian ini berhasil melakukan klasifikasi data SMS dengan menggunakan algoritma Naïve Bayes, Iterative Dychotomizer Version 3, dan Tree Augmented Naïve Bayes Classifier.
2. Dengan menggunakan pembagian data dimana digunakan 70% data *training* dan 30% data *testing* diperoleh bahwa algoritma Naïve Bayes Classifier menghasilkan akurasi yang paling tinggi dibandingkan dengan kedua algoritma yang lainnya sebesar 96.21%.
3. Dengan menggunakan pembagian data 80% data *training* dan 20% data *testing* diperoleh bahwa algoritma Iterative Dychotomizer Version 3 menghasilkan akurasi yang paling tinggi dibandingkan dengan kedua algoritma yang lainnya sebesar 96.47%.
4. Peneliti memberi saran untuk melakukan tahap *preprocessing* lanjutan yang dapat memperbaiki kata-kata yang tidak memiliki arti dalam bahasa Inggris, dan juga untuk menghapus kata-kata yang berisi nama *merk* ataupun produk barang tertentu. Untuk menghapus *merk*, penulis menyarankan untuk menambahkan khusus *merk* ataupun produk tertentu ke dalam *list stopwords*.

## DAFTAR PUSTAKA

- Aradea, S. A., Ariyan, Z., & Yuliana, A. (2011). Penerapan Decision Tree untuk Penentuan Pola Data Penerimaan Mahasiswa Baru. *Jurnal Penelitian Sitotika*, 7(1).
- Arifin, D. D., & Bijaksana, M. A. (2016). Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier. *2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 80–84. IEEE.
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29.
- Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied Computing and Informatics*.
- Dewi, I. N., & Supriyanto, C. (2013). Klasifikasi Teks Pesan Spam Menggunakan Algoritma Naïve Bayes. *Semantik*, 3(1).
- Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.
- Padmanaban, H. (2014). *Comparative analysis of Naive Bayes and tree augmented naïve Bayes models*.
- Sethi, G., & Bhootna, V. (2014). SMS spam filtering

application using Android. *Int. J. Comput. Sci. Inf. Technol*, 5(3), 4624–4626.

Srividhya, V., & Anitha, R. (2010). Evaluating preprocessing techniques in text categorization. *International Journal of Computer Science and Application*, 47(11), 49–51.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1–4), 43–52.