BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Ada banyak studi literatur yang membahas terkait dengan klasifikasi analisis sentimen. Sebelum memulai penelitian, perlu dilakukan kajian terhadap penelitian sebelumnya terkait klasifikasi analisis sentimen. Bahwari (2019) di dalam penelitiannya terkait analisis sentimen terhadap ulasan di media sosial Twitter tentang 6 perusahaan penerbangan. Algoritma *Random forest* digunakan untuk mengklasifikasi hasil kategorisasi pola analisis sentimen. Penerapan *sentiment analysis* dengan pendekatan pembelajaran mesin *Random Forest* menghasilkan nilai akurasi sebesar 75% (Bahwari, 2019). Neogi dkk (2021) juga menerapkan klasifikasi analisis sentimen pada 20.000 data tweet tentang protes petani. Klasifikasi sentimen ini menggunakan empat algoritma populer *(Naïve Bayes, Decision Tree, Random Forest, dan Support Vector Machine*). Hasil dari masing-masing metode ditemukan bahwa performa *Random Forest* memiliki akurasi tertinggi (Neogi dkk., 2021).

Klasifikasi analisis sentimen tidak hanya dilakukan pada platform sosial media, Shaheen dkk (2019) melakukan klasifikasi analisis sentimen terhadap data ulasan telepon seluler pada situs Amazon.com. Pendekatan analisis sentimen yang digunakan ialah pendekatan pembelajaran mesin dimana studi ini menerapkan tujuh algoritma klasifikasi. Dari hasil yang diperoleh, *Random forest* menjadi algoritma yang memiliki tingkat akurasi yang terbaik yakni 85% dari algoritma lainnya (Shaheen, 2019). Xu (2020) dalam studinya menyajikan peningkatan model klasifikasi sentimen dengan mengusulkan kerangka pembelajaran *continuous naïve bayes*. Data yang digunakan berasal dari domain yang berbeda yakni sentimen ulasan pada produk Amazon dan ulasan film. Hasil penelitian yang dilakukan menyatakan bahwa, model yang diusulkan dapat mengambil pengetahuan yang dipelajari dari domain sebelumnya untuk diterapkan pada domain yang baru. Selain

itu, model ini memiliki kapasitas yang lebih baik dalam menangani ulasan meskipun berasal dari domain yang berbeda (F. Xu dkk., 2020).

Penelitian terkait kelas data tidak seimbang juga telah banyak dibahas dalam berbagai literatur sebagai solusi dalam menghadapi masalah tersebut. Guo dkk (2017) mengungkapkan bahwa kasus kelas data tidak seimbang bukan hanya terjadi pada model klasifikasi umum (berbasis numerik) tetapi juga dapat terjadi pada analisis sentimen. Umumnya, kasus ini terjadi ketika satu kelas dalam kumpulan data memiliki jumlah lebih banyak (mayoritas) sehingga mendominasi kelas lainnya yang lebih sedikit (minoritas) (Guo dkk., 2017). Implementasi metode SMOTE dalam tugasnya menghadapi data tidak seimbang terbukti mampu meningkatkan akurasi pada algoritma klasifikasi tanpa terkecuali pada algoritma *Random Forest Classifier*.

Pada penelitiannya, Utari dkk (2020) menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) untuk menangani data tidak seimbang terhadap prediksi *drop-out* di fakultas ABC di Universitas XYZ. Berdasarkan studi yang terkait, metode SMOTE mampu memaksimalkan kinerja model dengan algoritma *random forest* sebagai metode klasifikasinya. Model RF + SMOTE pada prediksi *drop-out* menghasilkan tingkat akurasi sebesar 93,43% (Utari dkk., 2020). Mustaqim dkk (2019) di dalam penelitiannya juga menerapkan *Synthetic Minority Oversampling Technique* (SMOTE) untuk prediksi pemakaian alat kontrasepsi implant menggunakan algoritma *Neural Network Backpropagation*. SMOTE digunakan untuk menyeimbangkan data minor pada dataset implant serta mengurangi terjadinya *overfitting* (Mustaqim dkk., 2019).

Penerapan teknik kelas data tidak seimbang tidak hanya dibahas dalam studi pembelajaran berbasis numerik. Umer dkk (2021) di dalam penelitiannya membahas terkait penerapan teknik kelas data tidak seimbang pada penentuan sentimen kutipan (sitasi). *Synthetic Minority Oversampling Techniques* (SMOTE) digunakan sebagai pemecahan masalah kelas data tidak seimbang. Sebesar 98,26% tingkat akurasi yang diperoleh dari hasil kombinasi metode *Extra Tree Classifiers* dan pembobotan menggunakan TF-IDF pada set data seimbang menggunakan SMOTE (Umer dkk., 2021). Vinodhini dan Chandrasekaran (2017) mengusulkan

dan menganalisis dua teknik kelas tidak seimbang (*under-sampling* dan *oversampling*) pada model dengan algoritma ensemble SVM. Penerapan teknik yang digunakan berupa *Random under-sampling* (RUS) sebagai teknik *under-sampling* dan SMOTE sebagai teknik *oversampling*. Dari hasil analisis penelitian ini menyatakan bahwa, kedua teknik memiliki kinerja yang baik dari sisi besaran rasio. RUS memiliki kinerja yang lebih baik pada rasio ketidakseimbangan yang lebih rendah. Sebaliknya, SMOTE lebih baik pada rasio ketidakseimbangan yang lebih tinggi (Vinodhini dan Chandrasekaran, 2017).

Ali dkk (2021) mengusulkan beberapa teknik dalam penelitian mereka terkait analisis sentimen pendeteksi ujaran pada tweet berbahasa Urdu. Teknik tersebut diusulkan guna menghadapi tantangan analisis, salah satunya masalah kelas tidak seimbang. Untuk mengatasi masalah tersebut, penulis menggunakan SMOTE untuk menyeimbangkan kelas datanya. Hasil penelitian ini menyatakan bahwa, SMOTE mampu meningkatkan kinerja dua algoritma yang digunakan. Algoritma *Multinominal Naïve Bayes* memperoleh peningkatan kinerja sebesar 1,5%, sedangkan SVM memperoleh peningkatan sebesar 6% (Ali dkk., 2021).

2.2 Dasar Teori

2.2.1 Klasifikasi Sentimen

Sentiment analysis (SA) merupakan salah satu dari beberapa sub-bidang studi komputasi yakni *Natural Language Processing* (NLP). SA umumnya sebuah teknik dalam penambangan informasi berupa ekstraksi emosi berdasarkan opini atau sentimen (ulasan) (Liu, 2012). Banyak informasi yang dapat ditarik dari setiap penilaian atau analisis sentimen. Salah satu manfaat penambangan informasi bagi pemilik bisnis adalah dapat memperoleh informasi tentang tingkat keberhasilan produk yang baru dirilis serta menjadikan informasi tersebut sebagai sumber pengetahuan untuk menganalisis market pasar maupun karakteristik konsumen (Shaheen, 2019).

SA dikategorikan ke dalam tiga tingkatan yakni tingkat dokumen, tingkat kalimat, dan tingkat aspek. Analisis sentimen atau penambangan opini memiliki manfaat pada setiap tingkatan salah satunya pada tingkat dokumen. Pada tingkat

dokumen sendiri membahas mengenai bagaimana manampilkan sebuah sentimen pada keseluruhan dokumen menggunakan model yang tepat. Pada serangkaian proses SA, terdapat berbagai jenis variasi tugas di dalamnya. Satu diantaranya yakni *sentiment classification*. Secara teknis, klasifikasi sentimen melakukan pengelompokkan terhadap dua kelas yakni positif dan negatif. Data yang digunakan biasanya berupa data ulasan produk. Setiap ulasan produk memiliki peringkat berupa skor penilaian yang diberikan oleh pengulasnya (Liu, 2012; Saberi dan Saad, 2017).

2.2.2 *Text Pre-processing*

Text Preprocessing merupakan tahapan yang sangat penting dalam proses membangun sebuah model sentimen. Tahapan ini dilakukan untuk membersihkan dataset sentimen berupa teks yang terstruktur agar data siap untuk diolah. Selain itu, proses ini mampu meningkatkan kinerja dan akurasi model klasifikasi. Sebuah dataset ulasan yang berisi kata yang tidak expresif secara jelas, maka kata tersebut perlu dihapus (Shaheen, 2019). Berikut penjelasan terkait beberapa tahapan pada *preprocessing*.

1. *Case Folding*

   Case folding merupakan tahapan yang paling sederhana dan efektif di dalam *preprocessing* berbasis teks. Pekerjaan yang dilakukan pada tahapan ini yakni mengubah semua huruf dalam dokumen menjadi bentuk yang setara (*lowercase*). Beberapa karakter lainnya yang tidak termasuk dalam kategori huruf ‘a’ sampai ‘z’ akan dianggap *delimiter* yang harus dihilangkan. Karakter angka dan tanda baca seperti “[!”#$%&’()*+,-./:;<=>?@[\]^_`{|}~]” serta karakter kosong (*spacing*) juga perlu diperhatikan sebab tidak termasuk dalam kategori yang dibutuhkan (Nugroho, 2019).

2. *Tokenization*

   Tokenization adalah sebuah proses yang dilakukan untuk memisahkan setiap kata yang menyusun sebuah dokumen. Umumnya, setiap kata akan dipisahkan menjadi potongan-potongan kata, frasa, atau bagian yang bermakna yang biasa disebut sebagai token. Sebagai contoh, kalimat “Saya sangat suka wangi parfum ini”. Jika dilakukan tokenisasi maka kalimat tersebut akan dipisah menjadi

‘saya’ ‘sangat’ ‘suka’ ‘wangi’ ‘parfum’ ‘ini’. Proses ini juga dapat dilakukan pada paragraf maupun kalimat tergantung kebutuhan (Nugroho, 2019).

3. *Stopword Removal*

Setiap data teks yang dimiliki baik berupa kalimat atau paragraf perlu melalui tahapan filterisasi. Prinsip kerja *stopword removal* yakni dengan menghapus kata-kata yang rendah makna dan hanya kata-kata penting dari hasil token yang diambil. *Stopword* sendiri merupakan kata umum yang muncul namun tidak memiliki makna salah satunya seperti “yang”, ”dan”, ”di”, ”apa”, dll. *Stopword removal* merupakan salah satu proses penghapusan kata yang tidak bermakna (Nugroho, 2019). Penggunaan *stopword removal* juga sangat membantu dalam mengurangi dimensi pada data yang kita miliki.

4. *Text Normalization*

Seringkali ditemukan pada suatu ulasan konsumen terdapat penulisan kata slang. Kata slang (*slang words*) adalah sebuah kata yang tidak tergolong dalam standar kamus Indonesia (KBBI). Kata ini dapat berupa singkatan ataupun istilah-istilah gaul yang dipakai di masyarakat (Khomsah and Agus Sasmito Aribowo, 2020). Mengubah kata slang dengan melakukan normalisasi kata merupakan tahapan yang wajib dilakukan untuk mengubah kata-kata non standar ke kata yang sesuai dengan standar penulisan (Rahate dan Chandak, 2019). Sebagai contoh, “tdk” dikonversi “tidak”, “pengen” dikonversi menjadi “ingin”, “gue” menjadi “saya”, dan seterusnya. Maka dari itu, menurut Khomsah dkk (2020), para peneliti perlu membuat sebuah kamus tersendiri yang berisi kumpulan kata slang yang nantinya akan dikonvesi dan diproses.

2.2.3 *Term Frequency-Inverse Document Frequency (TF-IDF)*

Fungsi proses pembobotan kata adalah untuk memberikan bobot pada setiap fitur kata berdasarkan frekuensi kemunculan kata. Salah satu metode pembobotan yang terkenal yakni *Term Frequency-Inverse Document Frequency* (TF-IDF). *Term Frequency* (TF) merepresentasikan frekuensi kemunculan kata pada suatu dokumen. *Inverse Document Frequency* (IDF) membantu untuk mengetahui apakah *term* yang dicari telah sesuai dengan kata kunci yang diharapkan (Fitriyah, 2020). Metode TF-IDF dapat dihitung menggunakan persamaan berikut :

$$tfidf\ (t, d, D) = tf(t, d) * idf(t, D) \quad (2.1)$$

Di mana :

tf (*t*, *d*) = berapa kali *term* '*t*' muncul dalam dokumen '*d*'

idf (*t*, *D*) = *term 't'* dalam kumpulan dokumen keseluruhan '*D*'

Untuk memperoleh hasil dari persamaan '$idf(t, D)$', maka digunakan formula sebagai berikut :

$$idf(t, D) = log \frac{N}{|d \in D: t \in d|} \quad (2.2)$$

Di mana :

N = total dokumen yang ada di dalam corpus *D*

$|d \in D: t \in d|$ = jumlah dokumen dimana *term* '*t*' muncul.

2.2.4 Pembelajaran Tidak Seimbang

Pembelajaran tidak seimbang adalah suatu masalah klasifikasi yang sering muncul pada bidang data mining dan pembelajaran mesin berbasis pembelajaran terawasi. Pembelarajan tidak seimbang terjadi ketika dalam proses *pre-processing* data, ditemukan dua atau lebih kelas dimana satu sisi memiliki data yang sangat sedikit (*minority*), sementara sangat banyak (*majority*) pada data kelas lainnya (Elreedy and Atiya, 2019). Standarnya, suatu pembelajaran mesin dengan pendekatan klasifikasi diasumsikan untuk memiliki kelas dengan distribusi data yang seimbang. Sebab, jika terjadi masalah klasifikasi yang tidak seimbang, biasanya mesin akan lebih memprioritaskan distribusi data yang banyak dari pada data yang sedikit. Akibatnya, performa dari hasil pembelajaran cenderung baik dalam prediksi terkait data mayoritas daripada data minoritas (Soltanzadeh and Hashemzadeh, 2021).

Pada dasarnya, data atau kelas tidak seimbang mampu diatasi apabila data atau kelas baik mayoritas maupun minoritas memiliki jumlah distribusi yang seimbang. Penyamaan distribusi dapat dilakukan dengan metode *re-sampling* baik dengan membuat data sintetik dari kelas minoritas (*over-sampling*) atau membuang data dari kelas mayoritas asli (*under-sampling*) (Li dkk., 2018). Menurut Li dan Sun, dataset dikatakan tidak seimbang apabila proporsi sampel kelas minoritas kurang dari 35% dari dataset (Li and Sun, 2012). Sedangkan menurut Mustaqim,

suatu kelas data sudah dapat dikatakan seimbang apabila perbandingan data mayor dua kali ditambah satu dari kelas minornya (Mustaqim dkk., 2019).

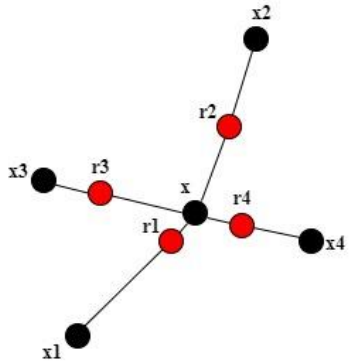
2.2.5 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE adalah salah satu teknik dalam klasifikasi pembelajaran yang digunakan untuk mengatasi masalah data atau kelas tidak seimbang (dalam hal ini terkait sampel minoritas). Teknik ini mampu menyeimbangkan antara data minor dengan data mayor serta mampu mengurangi masalah *overfitting* pada proses pembelajaran (Mustaqim dkk., 2019). Proses pembuatan data sintetis pada teknik ini yakni dengan melakukan interpolasi antara sampel minor dengan k-tetangga terdekat (k-Nearest Neighbor) hingga seimbang dengan sampel mayor (Soltanzadeh and Hashemzadeh, 2021). Berikut tahapan sederhana implementasi SMOTE pada sampel dataset tidak seimbang (Qu dkk., 2020) :

1. Memilih salah satu sampel x_i yang termasuk dalam sampel minoritas X.
2. Lalu menghitung jarak *Euclidean* antara x_i dengan sampel lainnya hingga mendapatkan *k* tetangga terdekat yang ditandai dengan persamaan sebagai y_i $(j = 1, 2, \dots, k)$.
3. Dalam pengambilan sampel data untuk x_i, sampel terdekat dipilih secara acak. Berikut formula yang digunakan agar tercipta data yang baru.

$$x_{new} = x_i + acak(0,1) \times \left(y_j - x_i\right) \quad (2.3)$$

Berikut visualisasi sampel data sintetik hasil dari SMOTE yang ditampilkan pada gambar 2.1 di mana data x_1, $x_2 \dots x_n$ adalah data minor acak sedangkan data r_1, r_2 ... r_n merupakan data sintetik baru.

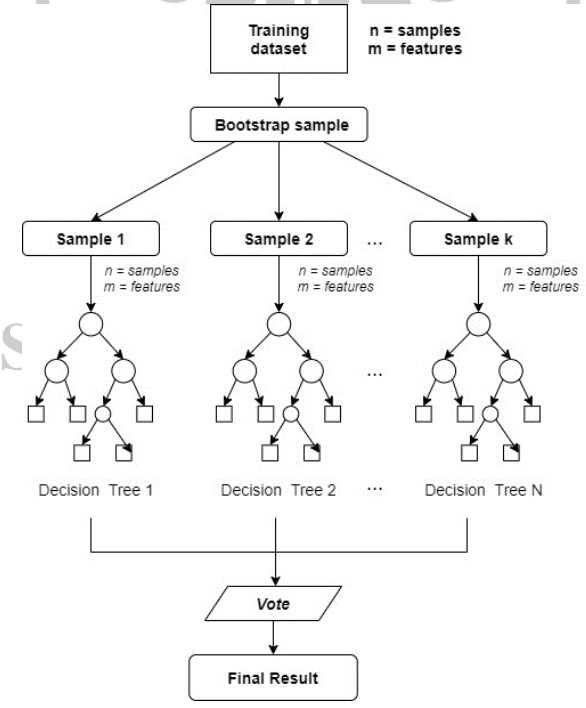


Gambar 2.1 Sampel Data Sintetik Hasil SMOTE

Salah satu kunci keberhasilan SMOTE dalam menghadapi data tidak seimbang ialah tergantung pada kemampuan algoritma ini mengikuti aturan distribusi sampel yang dekat dengan distribusi sebenarnya.

2.2.6 Algoritma Random Forest Classifier

Pada tahun 2001, *Random Forest* (RF) sebuah algoritma terawasi yang diusulkan oleh seorang Leo Breiman dari Universitas California (Breiman, 2001), Algoritma ini adalah hasil dari pengembangan klasifikasi dasar *Decision Tree*. Terdiri dari sejumlah pohon keputusan, algoritma RF mampu beroperasi secara acak dengan memilih subset sampel dan subset fitur dalam menjamin independensi pada pohon keputusan serta generalisasi yang lebih baik (Parmar dkk., 2019). RF sebagai algoritma lanjutan dari metode *bagging*, RF tidak memisahkan *node* dengan pemisah antar variabel tetapi melakukan pemilihan secara acak dari hasil beberapa node *decision tree* dan memilih pilihan terbanyak sebagai hasil prediksi, proses sebagaimana terlihat pada gambar 2.2. Selain itu, RF memiliki performa yang baik dalam menghadapi *overfitting* dan pengguna mampu membuat pohon keputusan sebanyak yang diinginkan (Bahwari, 2019)



Gambar 2.2 Ilustrasi *Random Forest Classifier*(Park dkk)

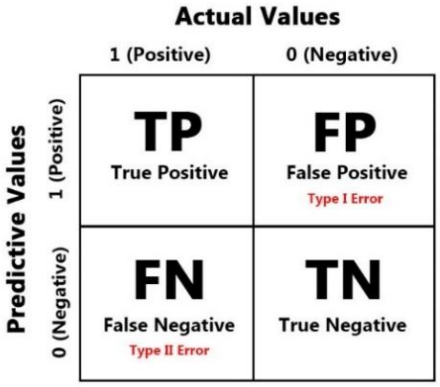
Pada algoritma RF, setiap hasil pemilihan *tree* akan sangat menentukan hasil prediksi yang dihasilkan pada proses klasifikasi. Semakin banyak *decision tree* dengan voting yang terbanyak, maka semakin mudah mesin dalam menentukan kelas target. Pembentukan sebuah hutan acak dapat dilakukan dengan prosedur diantaranya, dimulai dengan membuat gambaran sampel *bootstrap* (*Z*) yang berukuran *n* yang berasal dari data latih; melakukan pengulangan secara rekursif pada tahapan seleksi variabel *m*, lalu pengambilan atau pemisahan variabel terbaik diantara m, hingga membagi node ke dalam 2 bagian node; lalu menghasilkan *output* pohon ensemble $\{T\mathrm{b}\}_1^B$ (Breiman, 2001).

2.2.7 Metode Evaluasi

Untuk mengetahui tingkat performa suatu model ML, maka perlu dilakukan analisis model lebih lanjut dengan menggunakan metode tertentu. Pada penelitian ini, beberapa teknik evaluasi yang digunakan yakni teknik *confusion matrix* dan analisis ROC AUC.

1) *Confusion Matrix*

Confusion Matrix (CM) merupakan suatu alat dalam mengevaluasi model pada *supervised learning* yaitu algoritma klasifikasi (Bekkar dkk., 2013). CM menilai kinerja model klasifikasi berdasarkan jumlah fitur yang diprediksi dengan benar dan salah (Saifudin and Wahono, 2015). Memiliki kerangka kerja berbentuk matriks, model evaluasi ini memiliki dimensi 2×2 di mana kolomnya merupakan hasil kelas prediksi dan barisnya merupakan hasil kelas asli dalam kasus klasifikasi biner (J. Xu dkk., 2020). Berikut di bawah ini kerangka CM yang telah dijelaskan pada kalimat sebelumnya :



Gambar 2.3 *Confusion Matrix* Untuk Klasifikasi Dua Kelas

Pada gambar 2.3, terdapat akronim yang mewakili setiap hasil kelas pada proses klasifikasi antara lain TP, TN, FP dan FN. Untuk menentukan persamaan nilai akurasi, presisi, dan recall berikut rumus dasar nilai akurasi pada metode evaluasi matriks :

$$Akurasi = \frac{TP + TN}{TP+TN+FP+FN} \quad (2.4)$$

$$Sensitivitas = recall = TP_{rate} = \frac{TP}{TP+FN} \quad (2.5)$$

$$Spesifitas = TN_{rate} = \frac{TN}{TN+FP} \quad (2.6)$$

$$FP_{rate} = \frac{FP}{FP+TN} \quad (2.7)$$

$$Presisi = \frac{TP}{TP+FP} \quad (2.8)$$

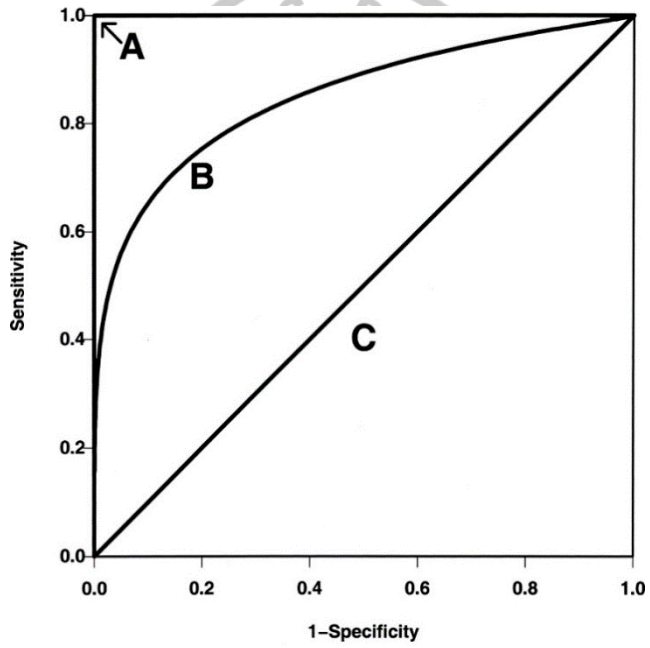
Berikut maksud dari beberapa akronim terkait :

- TP = *True Positive*, dimana data kasus positif yang diprediksi benar.
- TN = *True Negative*, dimana data kasus negatif yang diprediksi benar.
- FP = *False Positive*, dimana data kasus negatif namun diprediksi sebagai data positif.
- FN = *False Negatif*, dimana data kasus positif namun diprediksi sebagai data negatif.

2) Metode Evaluasi AUC ROC

Pada kasus kelas data tidak seimbang seringkali memperoleh hasil akurasi yang baik cenderung tinggi. Hal tersebut disebabkan model hanya fokus melihat kelas data mayoritas saja. Penggunaan metode AUC (*Area Under the ROC Curve*) dinilai tepat untuk kasus masalah tidak seimbang sebab AUC mampu mengevaluasi

prediktor secara komprehensif (Zhang and Wang, 2011) serta mampu menilai model mana yang lebih baik secara rata-rata. Evaluasi AUC (*Area Under the ROC Curve*) juga merupakan salah satu metode evaluasi yang populer digunakan pada klasifikai dengan kasus data tidak seimbang (Saifudin dan Wahono, 2015). ROC AUC merupakan instrument yang digunakan pula untuk menyajikan informasi mengenai kinerja algoritma klasifikasi dalam bentuk grafik kurva. Informasi kurva diperoleh berdasarkan hasil perhitungan *confusion matrix* yakni antara *False Positive Rate* (FPR) dengan *True Positive Rate* (TPR) (Zou dkk., 2007). Gambar 2.3 di bawah ini merupakan contoh bentuk penyajian hasil evaluasi menggunakan kurva ROC:



Gambar 2.4 Bentuk Penyajian Analisis Kurva ROC (Zou dkk., 2007)

Pada gambar di atas, terdapat tiga hipotesis kurva yang masing-masing merepresentasikan analisis akurasi. Simbol huruf ‘A’ mewakili area dengan hasil akurasi terbaik (di atas standar) yakni AUC = 1, ‘B’ merepresentasikan bentuk kurva ROC di mana AUC = 0.85, dan garis diagonal yang mewakili *random classifier* yang sesuai. Salah satu tanda adanya peningkatan pada hasil analisis adalah ketika AUC mendekati angka 1 atau kurva ROC bergerak ke arah simbol A. (Zou et al., 2007).