

IMPLEMENTASI MACHINE LEARNING DENGAN METODE TEXT MINING DAN ALGORITMA NAÏVE BAYES TERHADAP MASALAH PENGKLASIFIKASIAN KELUHAN PELANGGAN PADA TWITTER OPERATOR TELKOMSEL

Dannie Febrianto H

Program Studi Sistem Informasi, FTI, Institut Teknologi Budi Utomo Jakarta

danniefbrianto@gmail.com

Abstrak

Saat ini PT. Telkomsel, menggunakan peran media sosial sebagai bentuk kepedulian terhadap pelanggannya untuk menangani keluhan dari para pelanggannya. *Tweet* dari pelanggan Telkomsel pada sosial media Twitter ditangani oleh divisi *customer service* Telkomsel. Manualnya proses pengklasifikasian yang dilakukan oleh divisi *customer service* Telkomsel pada setiap narasi *tweet* keluhan “*complaint*” yang masuk ke Twitter @Telkomsel, membuat proses tersebut dinilai tidak efektif. Tujuan skripsi ini adalah untuk memberikan solusi terkait permasalahan pengklasifikasian *tweet* keluhan dan bukan keluhan dari para pelanggan Telkomsel, serta membuat *tools* yang dapat melakukan klasifikasi terhadap narasi *tweet* berbahasa Indonesia yang dibagi menjadi dua klasifikasi, yaitu: Keluhan dan Bukan Keluhan. Metode yang digunakan yaitu menggunakan Algoritma *Naïve Bayes* yang ditambahkan fitur SMOTE serta *Adaboost*. Disisi lain, *tools* Gataframework dan Rapidminer digunakan juga dalam penelitian ini untuk membantu *preprocessing* dan *cleansing* pada *dataset* hingga membantu menciptakan *corpus* dan *sentiment analysis*. Total *dataset* yang digunakan sebanyak 2200 data, setelah dilakukan *cleansing* dan *preprocessing* menjadi 2.000 data. Hasil dari metode yang diusulkan mendapat nilai *accuracy* 78,84%, *presicion* 79,92%, *rasio recall* 77,18% dan AUC 0.809%. Berdasarkan hasil dari metode yang diusulkan dapat disimpulkan terbukti efektif untuk pengklasifikasian narasi *tweet* keluhan dan bukan keluhan dari pelanggan Telkomsel.

Kata kunci: *Text Mining, Naïve Bayes, Metode SMOTE, Adaboost, Sentiment Analysis*

1. PENDAHULUAN

Twitter merupakan salah satu media sosial yang paling sering digunakan oleh PT Telekomunikasi Seluler (Telkomsel). dalam mengungkapkan pendapat atau keluhan yang dialami oleh pelanggan Telkomsel.

Keluhan secara definisi diartikan sebagai satu pernyataan atau ungkapan rasa kurang puas terhadap suatu produk atau layanan jasa, baik secara lisan ataupun tulisan dari penyampaian keluhan, baik internal maupun eksternal (Saputra et al., 2018).

Pendekatan pola klasifikasi penulisan secara elektronik pada sosial media twitter saat ini dapat menjadi objek penelitian yang menarik terkait dengan menggunakan metode *text mining*. *Text mining* adalah perluasan dari penggalian data untuk data tekstual (Dirgantara & Suryadarma, 2014).

Text mining pun melibatkan semua kegiatan dalam penemuan informasi dan data penting lainnya dari berbagai sumber-sumber tekstual (Hashimi et al., 2015)

PT Telekomunikasi Seluler, Tbk (Telkomsel), memerlukan sebuah *tools* yang

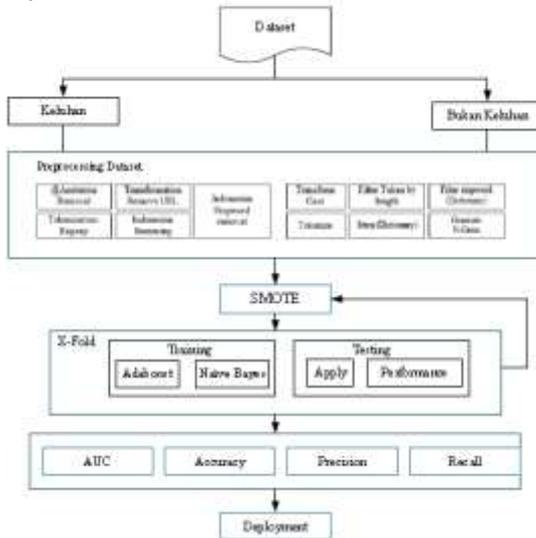
dapat melakukan ekstraksi terhadap narasi keluhan *tweet* berbahasa Indonesia yang dibagi menjadi dua kelas (Keluhan dan Bukan Keluhan). Berdasarkan hal tersebut metode *text mining* dapat membantu melakukan pengklasifikasian setiap narasi *tweet* keluhan.

2. METODOLOGI

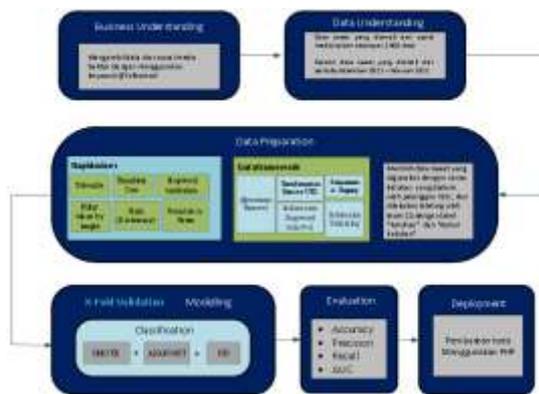
Penelitian yang akan dilakukan menggunakan data *tweet* para pelanggan Telkomsel di media sosial Twitter, dengan metode algoritma *Naïve Bayes* yang ditambahkan dengan *Adaboost* dan *feature Synthetic Minority Oversampling Technique Method* (SMOTE).

Model penelitian yang akan diusulkan dalam penelitian skripsi ini menggunakan model *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang telah disesuaikan dengan kebutuhan penelitian, metode *Cross-Industry Standard Process for Data Mining* (CRISP-DM) terdiri dari beberapa tahapan yaitu: *Business Understanding, Data Understanding, Data*

Preparation, Modelling, Evaluation, Deployment.



Gambar 1. Kerangka Pemikiran
Sumber : Penelitian Mandiri



Gambar 2. Model Penelitian CRISP-DM
Sumber : Penelitian Mandiri

3. HASIL DAN PEMBAHASAN

3.1 Business Understanding

Pemahaman mengenai objek penelitian dilakukan dengan mengambil data dari sosial media Twitter dengan menggunakan keyword @Telkomsel yang mengandung narasi keluhan dan opini atau pendapat dari para pelanggan Telkomsel.

3.2 Data Understanding

Tahap ini adalah proses memahami data yang akan digunakan sebagai bahan yang akan diteliti untuk bisa dilakukan ke tahap selanjutnya yaitu *Preprocessing*.

Menyiapkan data *tweet* yang diambil dari sosial media Twitter, dimana data tersebut sebanyak 2.200 data.

Proses pengklasifikasian dan *labelling* dilakukan oleh tim *customer services* Telkomsel.

3.3 Data Preparation

Tahap ini mempersiapkan data untuk melakukan langkah-langkah yang disebut dengan *text-preprocessing*, dengan menggunakan dua aplikasi *preprocessing* yaitu, Gataframework dan Rapidminer

3.4 Modelling Metode

Tahap ini *dataset* yang telah dibuat pada tahap sebelumnya digunakan sebagai masukan untuk algoritma klasifikasi, yaitu dijadikan sebagai *dataset training* dan *testing*.

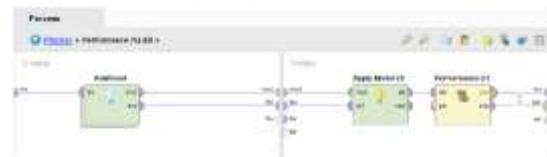
Pada penelitian ini akan digunakan satu jenis algoritma sekaligus untuk diuji, yaitu algoritma *Naïve Bayes* yang ditambahkan dengan *AdaBoost* dan *feature Synthetic Minority Oversampling Technique Method (SMOTE)*.

Teknik *preprocessing* menggunakan lima metode dari Gataframework yaitu, *@Anotation Removal*, *Transformation Remove URL*, *Tokenization (Regexp)*, *Indonesian Stemming*, *Indonesian Stopword Removal* dan *preprocessing* dari Rapidminer menggunakan enam metode yaitu, *Tokenization*, *Filter Tokens (by length)*, *stem (dictionary)*, *filter stopwords* tambahan, *generate N-gram* dan validasi *10 Cross fold validation*.

3.5 Evaluasi Hasil Pengujian dan Validasi Model

Hasil dari pengujian model yang dilakukan adalah mengklasifikasikan *tweet* yang mengandung narasi keluhan dan opini atau pendapat dari para pelanggan Telkomsel menjadi dua klasifikasi, yaitu, “Keluhan” dan “Bukan Keluhan” sebagai *dataset local* menggunakan algoritma *Naïve Bayes* yang ditambahkan dengan *AdaBoost* dan *feature Synthetic Minority Oversampling Technique Method (SMOTE)*.

3.5.1 Evaluasi Hasil Pengujian dan Validasi Model



Gambar 3. Desain Proses untuk NB + AdaBoost + SMOTE Klasifikasi Keluhan dan Bukan Keluhan

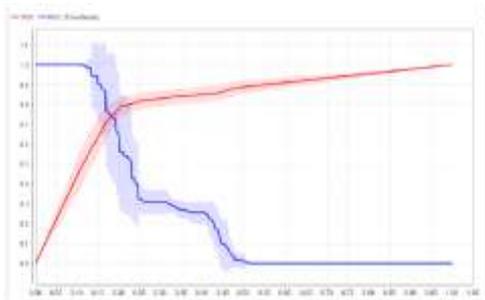
Sumber : Penelitian Mandiri

3.5.1.1 Confusion Matrix Algoritma NB + Adaboost + SMOTE

Tabel 1. Confusion Matrix NB + Adaboost + SMOTE

Sumber : Penelitian Mandiri

	True Bukan_keluhan	True Keluhan	Class Precision
Pred. Bukan_Keluhan	913	259	77,90%
Pred. Keluhan	221	875	79,84%
Class Recall	80,51%	77,16%	
Precision: 79,92% +/- 3,03% (micro average: 79,84%) (positive class: Keluhan)			
	True Bukan_keluhan	True Keluhan	Class Precision
Pred. Bukan_Keluhan	913	259	77,90%
Pred. Keluhan	221	875	79,84%
Class Recall	80,51%	77,16%	
Recall: 77,18% +/- 5,13% (micro average: 77,16%) (positive class: Keluhan)			
	True Bukan_keluhan	True Keluhan	Class Precision
Pred. Bukan_Keluhan	913	259	77,90%
Pred. Keluhan	221	875	79,84%
Class Recall	80,51%	77,16%	



Gambar 4. AUC Model NB + Adaboost + SMOTE

Sumber : Penelitian Mandiri

Berdasarkan hasil evaluasi dari proses pengujian finalisasi model algoritma *Naive Bayes* yang ditambahkan dengan *Adaboost* dan *feature* (SMOTE) yaitu mendapatkan hasil sebagai berikut: *Accuracy* 78,84%, *precision* 79,92%, *Recall* 77,18% dan *AUC* 0,809%.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai permasalahan pengklasifikasian *tweet* yang mengandung narasi keluhan dan opini atau pendapat dari para pelanggan PT Telekomunikasi Seluler,

Tbk., pada sosial media Twitter dengan menggunakan *mention keyword* @Telkomsel yang dibagi menjadi dua klasifikasi yaitu, “Keluhan” dan “Bukan Keluhan” dapat ditarik kesimpulan yaitu, pendekatan dengan menggunakan metode *text mining* dan pemodelan algoritma *Naive Bayes* yang ditambahkan dengan *Adaboost* dan *feature* (SMOTE), terbukti efektif dalam hal pengklasifikasian yang mengandung narasi *tweet* keluhan dan bukan keluhan dari para pelanggan Telkomsel.

Hasil penelitian ini juga menghasilkan model finalisasi dari klasifikasi pada hasil penelitian algoritma *Naive Bayes* yang ditambahkan dengan *Adaboost* dan *feature* (SMOTE) yaitu mendapatkan hasil sebagai berikut: *Accuracy* 78,84%, *Precision* 79,92%, *Recall* 77,18% dan *AUC* 0,809%.

DAFTAR PUSTAKA

Saputra, D. D., Pratama, B., Akbar, Y., & Gata, W. (2018). Penerapan Text Mining Untuk Assingment Complaint Handling Customer Terhadap Divisi Terkait Menggunakan Metode Decision Tree Algoritma C4.5 (Studi Case : Pt. XI Axiata, Tbk) Selection and peer-review under responsibility of The 11th STIKOM CKI on SPOT. *CKI On SPOT, 11*(2).

Dirgantara, U., & Suryadarma, M. (2014). Perhitungan Analisis Sentimen Berbasis Komparasi Algoritma Naive Bayes Dan K-Nearest Neighbour Berbasis Particle Swarm Optimization Pada Komentar Insiden Pembalap Motogp 2015. *Jurnal Sistem Informasi Universitas Suryadarma*. <https://doi.org/10.35968/jsi.v6i2>.

Hashimi, H., Hafez, A., & Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior, 51*, 729–733. <https://doi.org/10.1016/j.chb.2014.10.062>