SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK II - Surabaya, 26 Maret 2022
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2022.2720

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Klasifikasi *Sentiment Tweet* Pelanggan IndiHome Selama Pandemi Covid-19 Menggunakan Algoritma *Multinomial Naive Bayes*

Sigit Pamungkas[1], J.B. Budi Darmawan[2]

Informatika, Fakultas Sains dan Teknologi, Universitas Sanata Dharma[1,2]
e-mail: sigitpamungkas64.sp@gmail.com[1], b.darmawan@usd.ac.id[2]

ABSTRACT

During the Covid-19 pandemic, many school and office activities were carried out boldly. Requires an internet network to support these activities, so many people take the initiative to subscribe to the IndiHome internet service. Many customers are dissatisfied with IndiHome's services by writing tweets on Twitter. Sentiment analysis is needed on tweet comments for the IndiHome service application based on customer opinions. Classification using the Multinomial Naive Bayes method is applied to produce optimal accuracy. In the Naive Bayes Multinomial classification, the retrieval features are taken from a simple multinomial distribution. Sentiment analysis process consists of data labeling process using VADER Lexicon, pre-processing, then the number of term frequency and document frequency will be calculated. The data will be divided into training and testing data using k-fold Cross Validation and text classification using the Multinomial Naive Bayes method by finding the highest probability value. From the test results, the best average accuracy value is found in the 11-fold variation, which is 76.07%.

Keywords: *Multinomial Naïve Bayes, Classification, Sentiment analysis, VADER Lexicon*

ABSTRAK

Selama pandemi Covid-19 banyak kegiatan sekolah maupun kantor yang dilakukan secara daring. Diperlukan jaringan internet untuk mendukung kegiatan tersebut, maka banyak orang yang berinisiatif untuk berlangganan layanan internet IndiHome. Banyak pelanggan yang merasa tidak puas dengan pelayanan IndiHome dengan menuliskan *tweet* di Twitter. Diperlukan *sentiment analysis* pada komentar *tweet* untuk mengevaluasi layanan IndiHome berdasarkan opini pelanggan. Klasifikasi menggunakan metode *Multinomial Naive Bayes* diterapkan untuk menghasilkan akurasi yang optimal. Pada klasifikasi *Multinomial Naive Bayes* fitur diasumsikan diambil dari distribusi Multinomial sederhana. Proses analisis sentimen terdiri dari proses *labeling* data menggunakan *VADER Lexicon*, lalu *pre-processing*, selanjutnya

akan dihitung jumlah *term frequency* dan *documen frequency*. Data akan dibagi menjadi data *training* dan *testing* menggunakan *k-fold Cross Validation* serta klasifikasi teks dengan metode *Multinomial Naive Bayes* dengan mencari nilai probabilitas tertinggi. Dari hasil pengujian diperoleh nilai rata-rata akurasi terbaik terdapat pada variasi 11 *fold* yaitu sebesar 76,07%.

Kata kunci: *Multinomial Naive Bayes*, Klasifikasi, *Sentiment analysis*, *VADER Lexicon*

PENDAHULUAN

Selama pandemi Covid-19 layanan internet sangat dibutuhkan, karena banyak kegiatan sekolah maupun kantor yang dilakukan secara daring. Maka dari itu, dibutuhkan layanan jaringan internet WiFi yang lebih stabil dibandingkan data seluler untuk memudahkan pekerjaan dan kegiatan sekolah daring masyarakat. Penyedia jaringan internet WiFi IndiHome adalah salah satu penyedia layanan internet yang digunakan masyarakat Indonesia. Namun, dalam kurun waktu pandemi Covid-19, IndiHome sedang menerima banyak keluhan dari para pengguna mereka di media sosial khususnya Twitter. Tagar tentang keluhan jaringan Indihome seringkali menempati jajaran *Trending Topic* Twitter. Dikarenakan pelayanan *customer service* yang kurang memuaskan dan solutif, masyarakat memanfaatkan media twitter untuk memberikan komentar terhadap layanan internet Indihome yang bermasalah.

Diperlukan suatu *sentiment analysis* pada kicauan/*tweet* sebagai cara evaluasi layanan IndiHome. *Sentiment analysis* dapat menganalisa suatu sentimen, pendapat, penilaian, evaluasi, emosi dan sikap pada suatu entitas seperti jasa, produk, permasalahan, peristiwa, organisasi, topik, atau individu tertentu. Dalam penelitian ini menggunakan metode *Multinomial Naive Bayes*. Metode ini merupakan variasi lain dari metode klasifikasi *Naive Bayes*. Metode ini berlandasan bahwa antar atribut saling berkaitan satu dengan yang lain berdasarkan konteks kelas, dan mengabaikan semua ketergantungan antar atribut. Metode ini menggunakan sejumlah kecil data *training* sebagai estimasi untuk parameter, seperti variansi variabel dan rata-rata yang digunakan dalam klasifikasi.

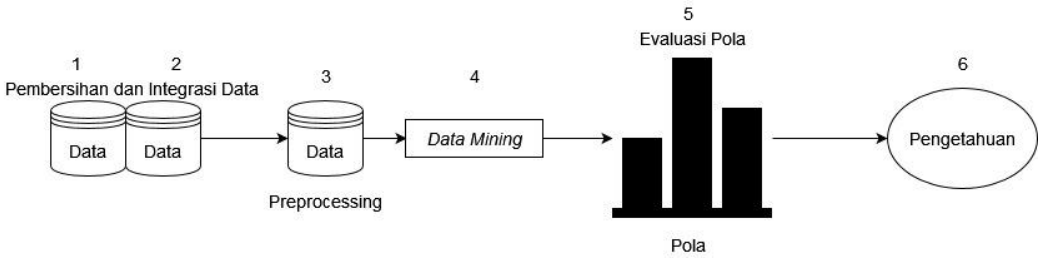
Penulis ingin melakukan penelitian klasifikasi sentimen menggunakan metode *Multinomial Naive Bayes* yang serupa dengan penilitian "Algoritma *Multinomial Naïve Bayes* Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter" (Yuyun dkk., 2021) dengan hasil akurasi sebesar 74% [1]. Sedangan penelitian "Analisis Sentimen Tentang Opini Pilkada DKI 2017 pada Dokumen Twitter Berbahasa Indonesia Menggunakan *Naive Bayes* dan Pembobotan *Emoji"* (Agnes Rossi Trisna Lestari, Rizal Setya Perdana, M. Ali Fauzi, 2017) akurasi yang didapat adalah 90% dengan rincian nilai *precission* 92%, *recall* 90% dan *f-measure* 90% [2]. Terbukti bahwa metode *Naive Bayes* memiliki performansi yang baik untuk melakukan klasifikasi *tweet* dilihat dari tingkat nilai akurasi di angka 90%. Penulis melakukan pelabelan data menggunakan *Library Vader Lexicon*, berbeda dengan penelitian "Analisis Sentimen Opini Film pada Twitter menggunakan metode *Naive Bayes*" oleh (Ratnawati, 2018) yang melakukan proses pelabelan data secara manual [3]. Pelabelan data menggunakan *Vader* sama dengan penelitian "Studi Analisis Metode Analisis Sentimen pada YouTube" (Santi Thomas dkk., 2021) [4]. Metode *Multinomial Naive Bayes* digunakan dalam penelitian ini tanpa memakai pembobotan pada emoji, penelitian ini berfokus dalam perhitungan probabilitas dari setiap kata.

METODE

Data Mining

Data mining merupakan penambangan atau penemuan informasi baru dengan aturan tertentu atau mencari pola dengan jumlah data yang begitu besar. *Data mining*, sering juga disebut sebagai KDD atau *knowledge discovery in database*. KDD merupakan kegiatan pemakaian, pengumpulan, historis data untuk menemukan hubungan atau keteraturan pola pada

dataset berukuran besar [5]. Tahap-tahap *data mining* disajikan pada gambar 1 di bawah ini. Mulai dari pengumpulan data sampai dengan diperoleh akurasi dari hasil klasifikasi oleh sistem.



Gambar 1 Tahap-tahap *Data Mining*

Tahap-tahap dalam alur proses sistem adalah:

1. Pembersihan data
   Dataset akan disaring dan hanya yang menandai akun @IndiHome saja yang digunakan untuk membatasi kata kunci pencariannya.
2. Integrasi data
   Data *tweet* akan ditranslate ke bahasa Inggris terlebih dahulu untuk memberikan label. Pelabelan data menggunakan library pada pyhton yaitu *VADER Lexicon.* Setelah data mempunyai label, maka akan diintegrasikan dengan data *tweet* awal yang berbahasa Indonesia.
3. *Pre-processing* Data
   Pre-processing ini terdiri dari beberapa tahapan untuk menjadikan kata pada *tweet* menjadi bahasa yang baku. Proses *preprocessing* terbagimenjadi *Case Folding, Tokenizing, Stopword Removal, Normalisasi, Stemming.*
4. Proses *Data Mining*
   Proses utama saat metode klasifikasi *Multinomial Naive Bayes* diterapkan untuk menemukan pengetahuan dari data penelitian.
5. Evaluasi pola
   Evaluasi untuk menilai apakah hipotesa yang ada memang tercapai dengan model prediksi maupun melalui pola-pola yang ada. Jika hasil tidak sesuai dengan hipotesa, maka hasil tersebut dapat dijadikan umpan balik untuk evaluasi *data mining.*
6. Presentasi pengetahuan
   Tahap ini merupakan presentasi bagaimana memformulasikan keputusan dari hasil analisis yang didapat.

Sentiment Analysis

Sentiment analysis adalah ekspresi secara tekstual dari hasil riset komputasional sentimen, emosi dan opini [6]. Tugas *sentiment analysis* adalah menentukan pendapat yang dari sebuah kalimat atau dokumen kemudian menganalisis apakah bersifat positif, negatif, atau netral [7]. Hasilnya berupa teks yang sudah dikelompokkan berdasarkan sentimen positif, negatif, atau netral dan dapat berupa *word plot atau* representasi visual dari data teks.

Pengumpulan Data

Data untuk penelitian berupa *tweet* yang menggunakan bahasa Indonesia yang diperoleh dari media sosial Twitter. *Tweet* yang menandai akun @IndiHome digunakan sebagai kunci dengan batasan waktu selama Pandemi Covid-19 pada tanggal 1 sampai 9 Desember 2020. Penulis membatasi *tweet* yang akan diambil dengan filter @IndiHome

berjumlah 2000 *tweet*. Dengan akses API dari Twitter proses *crawling* data dilakukan dengan menggunakan bahasa pemrograman R lalu data akan disimpan ke dalam bentuk excel. Pelabelan yang digunakan hanya negatif dan positif saja yang dipakai. Label netral tidak dipakai karena tidak dapat dijadikan sebagai bahan evaluasi untuk meningkatkan layanan IndiHome. Dari 2000 *tweet*, setelah *preprocessing* data menjadi 1960 *tweet.*

Vader Lexicon

Vader merupakan singkatan dari *Valence Aware Dictionary for Social Reasoning*, diperkenalkan pada tahun 2014 oleh C.J Hutto dan Eric Gilbert. Pembentukan metodenya berdasarkan pendekatan *human-centric*, dengan penggabungan antara validasi empiris dan analisis kualitatif menggunakan penilaian dan kebijaksanaan manusia [8]. *Vader* digunakan sebagai model analisis sentimen yang dapat menentukan keragaman data sesuai dengan intensitas kekuatan emosional yang tersedia pada kamus data *Lexicon* [9]. Metode *vader lexicon* adalah sebuah *package* yang tersedia pada bahasa pemrograman *phyton* dari fitur NLTK (*Natural Language Toolkit*). Kamus *lexicon* berfungsi untuk menentukan nilai kalimat, sentimen dan frasa. Sentimen dapat dikasifikasikan menjadi negatif, positif dan netral atau dapat berupa numerik seperti skor atau kisaran intensitas [10]. Sebelum dilabeli menggunakan kamus *Vader Lexicon*, data terlebih dahulu ditranslate ke dalam bahasa Inggris.

Tabel 1. Pelabelan Menggunakan *Vader*

No.	Tweet	Compound	Label
1.	since yesterday the internet cannot be accessed and always error, it's just expensive	{'neg': 0.184, 'neu': 0.816, 'pos': 0.0, 'compound': -0.4019}	NEGATIF
2.	want to ask, how to change the ordinary wps into wps lite?	{'neg': 0.0, 'neu': 0.894, 'pos': 0.106, 'compound': 0.0772}	POSITIF

Multinomial Naive Bayes

Multinomial Naive Bayes merupakan suatu metode *conditional probability* tanpa memperhitungkan informasi maupun urutan kata pada suatu kalimat atau dokumen. Metode ini hanya memperhitungkan banyaknya kata saja yang mucul dalam suatu dokumen. Contohnya terdapat dokumen *d* dan himpunan kelas *c* [11]. Untuk memperhitungkan kelas dari dokumen d, maka dapat dihitung dengan persamaan(1).

$$P(c|term\ dokumen\ d) = P(c)\ .P(t_1|c).P(t_2|c)\ .P(t_3|c)\ .\ldots\ P(t_n|c) \quad \ldots\ldots\ldots(1)$$

Keterangan :

$P(c)$ = Probabilitas *prior* dari kelas

$P(c|term\ dokumen\ d)$= Probabilitas suatu dokumen termasuk kelas c

$P(t_n|c)$= Probabilitas kata ke- n dengan diketahui kelas c

Probabilitas prior kelas c ditentukan dengan persamaan (2).

$$P(c) = \frac{N_c}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

Keterangan :

N_c= Jumlah kelas c pada seluruh dokumen

N = Jumlah seluruh dokumen

Probabilitas kata ke- n dapat dihitung dengan menggunakan teknik *laplacian smoothing* disajikan dalam persamaan (3).

$$P(t_n|c) = \frac{count\,(t_{n,}c)+1}{count\,(c)+|V|} \quad \text{..... (3)}$$

Keterangan :

$count\,(t_{n,}, c)$ = Jumlah term $t_{n,}$yang ditemukan di seluruh data pelatihan dengan kategori c

$count\,(c)$= Jumlah term di seluruh data pelatihan dengan kategori c

V = Jumlah seluruh tem pada data pelatihan

HASIL DAN PEMBAHASAN

Pengujian menggunakan data 1960 *tweet* hasil *preprocessing* dengan variasi *k-fold* sebagai parameter, dengan variasi nilai *k* (7, 9,11). Variasi pengujian menggunakan nilai *k* (7, 9, 11) dilakukan untuk mengetahui hasil perbandingan dari setiap *fold* yang diuji. Hasil pengujian menggunakan *k* (7, 9, 11) disajikan pada tabel 2 di bawah ini.

Tabel 2. Hasil Pengujian variasi 7, 9 dan 11 *fold*

	***Total True* Negatif**	***Total True* Positif**	***Total False* Negatif**	***Total False* Positif**	***Rata-rata Precision***	***Rata-rata Recall***	***Rata-rata* Akurasi**	***Rata-rata F1-score***
7 *fold*	423	1061	340	137	88,62%	75,7%	75,71%	81,63%.
9 *fold*	431	1058	332	139	88,33%	76,06% .	75,97%.	81,71%.
11 *fold*	435	1056	328	141	88,21%.	76,28% .	76,06%	81,78%.

Percobaan menggunakan variasi 11 *fold* menghasilkan rata-rata akurasi tertinggi sebesar 76,7%. Nilai rata-rata *precision* dari variasi 11 fold sebesar 88,21%. Nilai rata-rata *recall* dari variasi 11 fold sebesar 76,28%. Nilai *F1-score* dari variasi 11 fold diperoleh sebesar 81,78%. Pada variasi 11 fold, sebanyak 435 tweet diklasifikasikan negatif dengan benar, 1056 *tweet* diklasifikasikan positif dengan benar. Sementara sebanyak 328 tidak relevant dengan label awal negatif dan 141 tweet tidak relevant dengan label awal positif.

KESIMPULAN

Pada penelitian ini digunakan dataset dari Twitter yang menandai akun @IndiHome selama pandemi Covid-19 sebanyak 1960 *tweet* terdiri dari 793 *tweet* negatif dan 1167 *tweet* positif. Penelitian klasifikasi sentimen ini menggunakan algoritma *Multinomial Naive Bayes* dengan variasi *K-Fold Cross Validation* 7, 9 dan 11. Berdasarkan hasil pengujian diperoleh model dengan akurasi tertinggi sebesar 76.06% dari variasi pengujian *11 fold*.

DAFTAR PUSTAKA

[1] Yuyun, Hidayah Nurul, Supriadi. (2021) 'Algoritma *Multinomial Naïve Bayes* Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter' *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, Vol. 5 No. 4 (2021) 820 - 826. doi: 10.29207/resti.v5i4.3146.

[2] Nurjanah, W. E., Perdana, R. S. and Fauzi, M. A. (2017) 'Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet', *Jurnal Pengembangan*

Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 1(12), pp. 1750–1757.

[3] Ratnawati, F. (2018) 'Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter', *JIFOTECH (Journal of Information Technology)*, Vol. 1 No.1(2021).

[4] Santi Thomas, Yuliana, Noviyanti. P (2021) 'Studi Analisis Metode Analisis Sentimen pada YouTube', *INOVTEK Polbeng - Seri Informatika*, 3(1), p. 50. doi: 10.35314/isi.v3i1.335.

[5] Romadloni, N. T., Santoso, I. and Budilaksono, S. (2019) 'Perbandingan Metode Naive Bayes , Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl', *Jurnal IKRA-ITH Informatika*, 3(2), pp. 1–9.

[6] Ira Zulfa and Edi Winarko. Sentimen analisis tweet berbahasa indonesia dengan deep belief network. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 11:187, 07 2017.

[7] Dehhaf (2010)Sentiment Analysis, Hard But Worth It!. [Online]. (update, 10 Maret 2010) Available at: http://customerthink.com/sentiment_ analysis_hard_but_worth_it/

[8] Hutto, C. J., & Gilbert, E. E. (2014). *VADER*: A Parsimonious Rule-based Model for *Sentiment analysis* of Social Media *Text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14).". *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*

[9] Elbagir, S., & Yang, J. (2019). Twitter *sentiment analysis* using natural language toolkit and *Vader sentiment*. *Lecture Notes in Engineering and Computer Science*, *2239*, 12–16.

[10] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Con*text*ual semantics for*sentiment analysis* of Twitter. *Information Processing and Management52*(1), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

[11] Destuardi dan Surya, S. 2009. "Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes". Surabaya: Teknik Elektro, Institut Teknologi Sepuluh November.