Jurnal Teknologi Informasi dan Komunikasi
Vol: 13 No 1 2022 **E-ISSN: 2477-3255**
Diterima Redaksi: 17-01-2022 | Revisi: 13-03-2022 | Diterbitkan: 25-04-2022

The Impact of Feature Extraction to Naïve Bayes Based Sentiment Analysis on Review Dataset of Indihome Services

Salsabila Mazya Permataning Tyas[1], Bagus Setya Rintyarna[2], Wiwik Suharso[3]
[1,3]Program Studi Teknik Infromatika, Fakultas Teknik, Universitas Muhammadiyah Jember
[2]Program Studi Teknik Elektro, Fakultas Teknik, Universitas Muhammadiyah Jember
[1,2]Jl. Karimata No. 49 Jember
e-mail: [1]salsa@unmuhjember.ac.id, [2]bagus.setya@bagus.setya.ac.id,
[3]wiwik@unmuhjember.ac.id

Abstract

Indihome is a product of PT Telekomunikasi Indonesia as an internet service provider or internet service provider (ISP) in Indonesia. Every product or service offered to the public certainly has its advantages and disadvantages, as well as Indihome. From the advantages and disadvantages of Indihome services, we can do a technique, namely sentiment analysis. In this study, sentiment analysis was carried out regarding public responses or reviews about IndiHome services on Twitter social media. This study uses a comparison of TF-IDF and Word2Vec feature extraction, and the classification method used is the nave Bayes classifier. The accuracy results obtained in this study were 96% using the TF-IDF feature extraction and testing was carried out using an unseen data test that was selected randomly resulting in an accuracy of 92%. While the accuracy value obtained by using the Word2Vec feature extraction is 60% by testing using unseen test data that was selected randomly resulting in an accuracy value of 44%.

Keywords: *Sentiment Analysis; Indihome; TF-IDF; Word2Vec; Naïve Bayes*

Pengaruh Ekstraksi Fitur terhadap Analisis Sentimen pada Data Review Pelayanan Indihome Berbasis Naïve Bayes

Abstrak

Indihome merupakan salah satu produk dari PT Telekomunikasi Indonesia sebagai penyedia jasa layanan internet atau internet servive provider (ISP) yang ada di Indonesia. Setiap produk atau jasa yang ditawarkan kepada masyarakat tentunya memiliki kelebihan dan kekurangannya masing-masing, begitu pula dengan Indihome. Dari kelebihan dan kekurangan tentang pelayanan Indihome dapat kita lakukan suatu teknik yakni sentimen analisis. Pada penelitian ini dilakukan analisis sentimen mengenai tanggapan atau review masyarakat tentang pelayanan IndiHome di media sosial twitterTeknik yang digunakan untuk analisis sentiment pada data review pelayanan indihome adalah Naïve Bayes Classifier. Fokus penelitian ini adalah untuk membandingkan kinerja ekstraksi fitur TF-IDF dan Word2Vec. Metrics yang digunakan untuk mengukur kinerja kedua teknik seleksi fitur adalah akurasi. Hasil eksperimen

dengan k-fold cross validation mengkonfirmasi bahwa TF-IDF lebih baik dibandingkan Word2Vec.

Kata kunci: *Analisis sentimen; Indihome; TF-IDF; Word2Vec; Naïve Bayes*

1. Pendahuluan

Sebagai produk dari PT Telekomunikasi Indonesia, salah satu perusahaan telekomunikasi terbesar di Indonesia, Indihome adalah penyedia layanan *Internet Service Provider* yang banyak digunakan masyarakat Indonesia. Tanggapan atau umpan balik yang diberikan oleh masyarakat tentang pelayanan indihome ini merupakan aset penting bagi perusahaaan untuk menentukan kualitas produk [1] yang mereka tawarkan.

Sentiment Analysis merupakan proses komputasi untuk mengeskstrak, memahami, dan juga mengolah suatu data yang tidak terstruktur menjadi data terstruktur. Dengan tujuan untuk menghasilkan suatu informasi sentimen yang terdapat pada pendapat, emosi, atau komentar seseorang dari dataset yang tidak terstruktur [2]. *Sentiment analysis* memberikan dampak dan manfaat yang sangat besar dalam berbagai bidang, sehingga menyebabkan *sentiment analysis* ini berkembang dengan pesat dan banyak digunakan.

Dalam melakukan analisis sentiment terdapat tahap ekstraksi fitur [3] (*fitur extraction*). Ekstraksi fitur merupakan suatu tahap untuk memproses sebuah kata yang menjelaskan suatu sentimen yang terdapat pada *dataset* untuk diekstraksi menjadi sebuah fitur atau aspek [2]. Salah satu fitur yang paling sering digunakan ialah TF-IDF (*Term Frequency-Inverse Document Frequency*). Fitur TF-IDF mempunyai keunggulan dibandingkan dengan fitur lainnya yaitu pengimplementasinya yang mudah. Akan tetapi, kekurangan dari fitur TF-IDF ini adalah tidak dapat memproses relasi semantik antar kata sehingga fitur ini menganggap setiap kata memiliki konteks yang berbeda [3]. Fitur yang menggunakan relasi semantik dapat dilakukan dengan metode *Word2Vec*. *Word2Vec* merupakan suatu metode yang merepresentasikan suatu kata pada ruang vektor dengan dimensi yang tinggi [4]. Selain itu, metode *word2vec* juga dapat mengetahui hubungan semantik antar kata dengan cara menghitung *cosine similarity* antar kata melalui nilai vektor [4]. Setelah dilakukan ekstraksi fitur, maka akan memasuki proses klasifikasi sentimen.

Tujuan penelitian ini adalah untuk mengetahui pendapat masyarakat terhadap pelayanan indihome berdasarkan data twitter dengan mengimplementasikan metode klasifikasi Naïve Bayes Classifier dan melakukan perbandingan ekstraksi fitur antara TF-IDF dan Word2Vec. Metode *Naïve Bayes Classifier* ialah jenis metode klasifikasi dengan menggunakan algoritma sederhana akan tetapi mempunyai kecepatan serta nilai akurasi tinggi [5]. Metode *Naïve Bayes Classifier* adalah suatu metode yang berdasar pada teori *Bayes* dengan asumsi yang kuat, dengan efek nilai atribut tidak bergantung pada kelas dan juga atribut lain. Teori *Bayes* merupakan suatu teori yang membahas klasifikasi untuk meramalkan atribut kelas suatu anggota probabilitas menurut data yang disediakan [6].

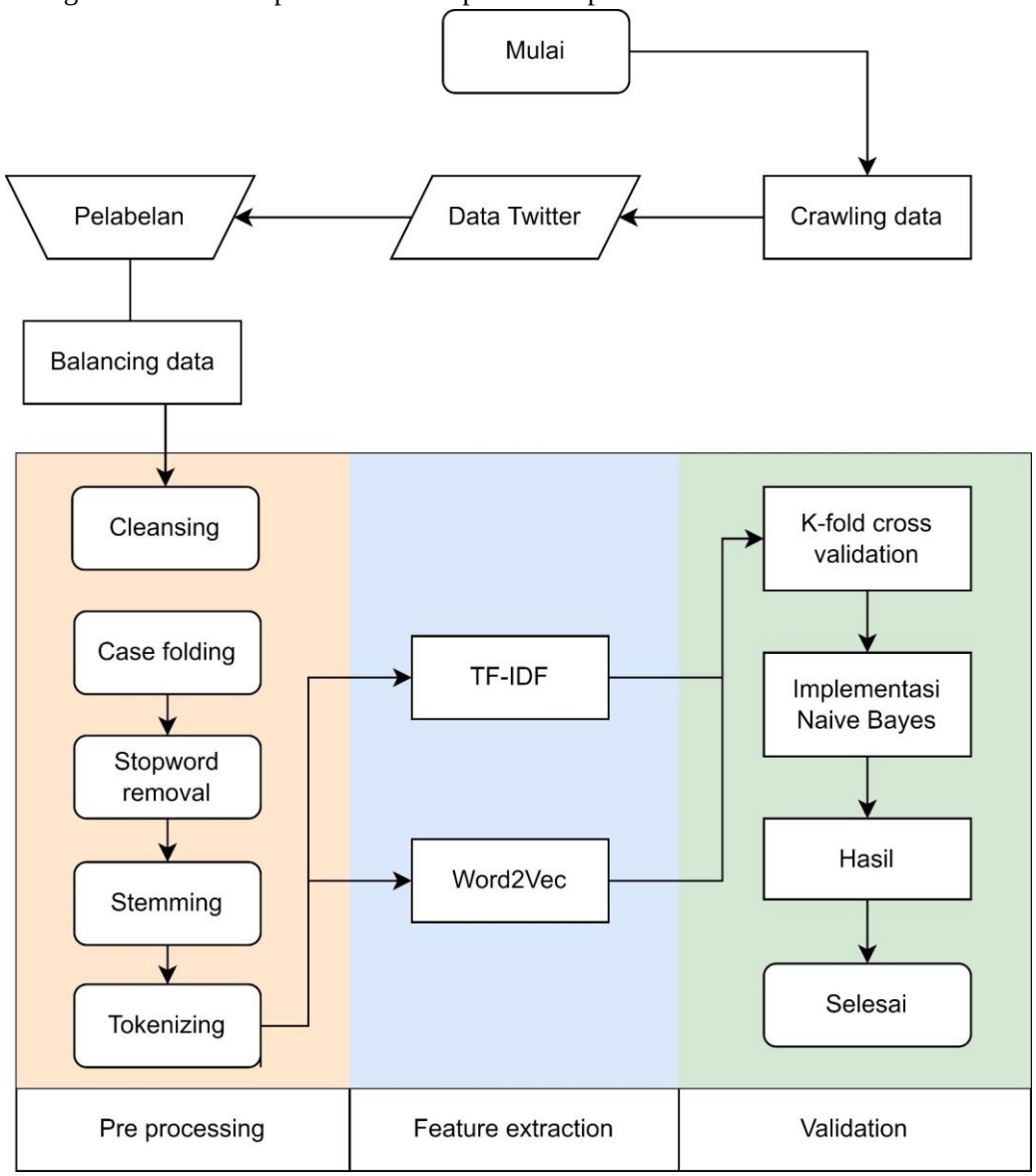
Penelitian sebelumnya yang dilakukan oleh Beakcheol Jang (2019) [7], telah dilakukan penelitian yang berjudul *Word2Vec Convolutional Neural Networks For Classification Of News Articles And Tweets*. Pada penelitian tersebut menjelaskan bahwa penerapan *word2vec* yang membahas tentang hubungan semantik antar kata secara signifikan mampu meningkatkan kinerja model klasifikasi [8]. Selain itu, penelitian sebelumnya yang dilakukan oleh Susanti Fransiska (2020) [9] dengan judul penelitian *Sentiment Analysis of Government Policy on Corona Case Using Naïve Bayes Algorithm*. Pada penelitian tersebut menggunakan ekstraksi fitur TF-IDF yang menghasilkan nilai akurasi sebesar 81% dengan nilai presisi sebesar 78%, *Recall* sebesar 91%, dan juga *f1-Score* sebesar 84%. Penelitian lain juga menggunakan algoritma Naïve Bayes dan menguji teknik ekstraksi fitur [10]. Namun dalam penelitian tersebut teknik ekstraksi fitur yang diuji adalah teknik seleksi fitur konvensional berbasis Gaussian dan Kernel Density. Berdasarkan tujuan dan beberapa referensi kunci pada penelitian pendahuluan, kebaruan yang disajikan dalam penelitian ini mencakup penerapan Word2Vec sebagai teknik

https://doi.org/10.31849/digitalzone.v13i1.9158

ekstraksi fitur terbaru untuk jenis data teks. Penelitian ini mengevaluasi kinerja Word2Vec dibandingkan teknik TF-IDF yang sederhana berbasis bag of word pada skenario sentiment analysis berbasis Naïve Bayes.

2. Metode Penelitian

Penelitian tentang pengaruh ekstraksi fitur terhadap analisis sentiment pada data *review* pelayanan indihome berbasis *naïve bayes* terdapat langkah-langkah penting yang hars dilakukan yaitu dimulai dengan pengambilan data dengan teknik *crawling*, kemudian dilakukan pelabelan, *balancing* data, lalu diproses di tahapan *preprocessing*, pembobotan kata TF-IDF dan *Word2Vec*, setelah itu melalui tahap validasi dan implementasi *naïve bayes*. Secara umum langkah-langkah dari metode penelitian ini dapat dilihat pada Gambar 1 di bawah ini.



Gambar 1. Tahapan Penelitian

2.1. Pengumpulan Data

Data yang diambil berasal dari twitter, pengambilan data tersebut dilakukan dengan menggunakan Twitter API (*Application Programming Interface*). Proses pengambilan data ini menerapkan bahasa pemrograman *python* dan memanfaatkan *library tweepy* yang telah tersedia.

Data yang dikumpulkan dari proses crawling ini menggunakan kata kunci indihome dengan jumlah data sebanyak 655 data pada bulan Juli 2021.

2.2. Pelabelan Data

Data pada penelitian ini dilakukan pelabelan secara manual. Pemberian label dilakukan oleh ahli bahasa yang berkompeten dalam bidangnya.

2.3. *Balancing* Data

Ketidakseimbangan jumlah proposi data pada penelitian ini diatasi dengan menggunakan teknik *sampling*. Teknik sampling ada beberapa macam yaitu *oversampling, undersampling,* dan *bothsampling* [11]. Namun, pada penelitian ini akan dilakukan dengan menggunakan teknik *undersampling*. Teknik *undersampling* merupakan suatu proses sampling yang dilakukan dengan cara mengurangi atau mengeliminasi sebagian data pada kelas mayoritas pada data. Proses eliminasi tersebut dapat dilakukan secara random (paling sederhana) sehingga biasa disebut dengan *random undersampling.* Selain itu, *undersampling* juga dapat dilakukan dengan menggunakan perhitungan statistik yang biasa disebut dengan *informed undersampling* [9].

2.4. *Preprocessing Text*

Preprocessing merupakan langkah guna menyiapkan data tekstual yang akan dipakai supaya dapat diproses pada langkah selanjutnya. Preprocessing terdiri dari beberapa tahap. Tahap-tahap yang dilakukan ialah sebagai berikut:

a. Cleansing

Cleansing data merupakan tahapan yang ditujukan untuk pembersihan data *tweet* dari kata-kata dan simbol yang tidak dibutuhkan. Tujuan *cleansing data* yaitu untuk mengurangi *noise* pada proses klasifikasi. Data *tweet* yang diperoleh dari proses crawling data dengan kata kunci indihome mengandung beberapa komponen atau karakteristik seperti *hashtag* (#), RT, username twitter (@), dan alamat situs (URL). Komponen tersebut tidak berpengaruh penting, maka perlu dibersihkan.

b. Case folding

Case folding merupakan langkah-langkah yang dilakukan untuk mengubah teks yang ada di dalam dokumen menjadi *lower case* (huruf kecil). Tujuan dilakukan tahapan ini supaya penulisan semua huruf pada teks menjadi sama atau seragam.

c. Stopword Removal

Stopword removal ialah sebuah tahapan untuk menghilangkan kata yang dianggap tidak memiliki arti atau tidak berpengaruh. Kata-kata tersebut terdapat pada kamus data *stopword* yang secara umum digunakan (terdiri dari yang, di, ke, dari, saya, dan lain-lain).

d. Stemming

Stemming ialah sebuah tahapan untuk mengubah kata berimbuhan menjadi kata dasar.

e. Tokenization

Tokenization merupakan tahapan untuk memecah teks menjadi satuan kata. Tujuan dilakukan tahapan *tokenization* ini untuk memisahkan setiap kata yang dapat diperlakukan sebagai pemisah atau bukan. Contoh pemisah kata yaitu spasi, enter, dan tabulasi.

2.5. Ekstraksi Fitur

Proses klasifikasi sentimen, pemilihan fitur merupakan hal yang penting untuk diperhatikan. Penelitian ini menggunakan perbandingan seleksi fitur antara TF-IDF dan *Word2Vec* untuk pembobotan data.

a. TF-IDF

Rumus dari TF-IDF adalah sebagai berikut:

$$tf_{t,d} = \frac{n_{t,d}}{T_{t,d}} \quad (1)$$

$$IDF = \frac{T_d}{f_{d,t}} \quad (2)$$

$$W_{d,t} = tf_{d,t}\ x\ idf_{f,t} \quad (3)$$

https://doi.org/10.31849/digitalzone.v13i1.9158

$$W_{d,t} = tf_{d,t}\ x \log\frac{D}{d_f} \quad (4)$$

Di mana $tf_{t,d}$ adalah frekuensi term, d_f adalah frekuensi dokumen yang mengandung term, D adalah jumlah total dokumen dan $W_{d,t}$ adalah bobot TF-IDF.

b. Word2Vec

Ekstraksi fitur *word2vec* [12] akan membuat model menggunakan *library gensim*. Pada proses training, dilakukan pengubahan data menjadi bentuk *one-hot-encoding*. Berikut ini merupakan langkah-langkah perhitungan *word2vec* [13]:

1. Persiapan Data

Pada tahapan ini dilakukan *one-hot vector encoding*, yaitu mengubah kata target menjadi 1 dan lainnya menjadi 0.

2. Calculate Hiden layer dan y_pred

Proses pembobotan pada *word2ve*c berada pada *hidden layer*, terdapat dua pembobotan yaitu W1 dan W2. Pada default *word2vec*, kedua bobot tersebut diinisialisasikan sebagai *random number* sesuai jumlah m x d. Namun pada perhitungan ini W1 dan W2 diinisialisasikan antara -1 sampai 5 sebagai berikut: 1) Pertama yaitu mencari h atau *hidden layer* dari suatu wt, 2) **Mencari *Error* dan *Sum of Different*, 3)** Selanjutnya mencari jumlah *SUM of Different* pada wt atau $\sum I$. Sedangkan untuk mencari error sendiri diperoleh darihasil pengurangan *one-hot encoding* dengan *softmax, 4)* ***Backpropagation***, Simpelnya proses ini bermaksud untuk menyesuaikan kembali tiap *weight* dan bias berdasarkan *error* yang didapatkan. Pada tahap ini akan dicari nilai delta dari masing-masing W. Untuk delta pada W2 didapatkan dengan perkalian h dan $\sum I$, dan 5) ***Update Weight,*** setelah diperoleh nilai delta dari masing-masing *weight*, selanjutnya meng-update nilai dari masing- masing *weight* dengan cara *Weight – learning rate* (0.025) * *Delta Weight*.

2.6. *Cross Validation*

Pada proses cross fold validation data akan dibagi menjadi k bagian sama banyak. Fungsi dari k-*fold cross validation* [14] yaitu menentukan jumlah data testing dan memisahkan data. Pada penelitian ini menggunakan 4 macam k-*fold* [15] yaitu:

1. 2-*fold cross validation*

2-*fold cross validation* memiliki maksud bahwa data akan dibagi menjadi 2 bagian sama banyak sehingga masing-masing bagian berjumlah 150 data. Fungsi dari fold 2 yaitu untuk menentukan jumlah data testing yaitu:

$$\frac{banyak\ data}{k\ fold} = \frac{300}{2} = 150$$

Sehingga 150 data digunakan sebagai data testing dan jumlah data sisanya digunakan sebagai data training.

2. 3-*fold cross validation*

3-*fold cross validation* memiliki maksud bahwa data akan dibagi menjadi 3 bagian sama banyak sehingga masing-masing bagian berjumlah 100 data. Fungsi dari fold 3 yaitu untuk menentukan jumlah data testing yaitu:

$$\frac{banyak\ data}{k\ fold} = \frac{300}{3} = 100$$

Sehingga 100 data digunakan sebagai data testing dan jumlah data sisanya digunakan sebagai data training.

3. 5-*fold cross validation*

5-*fold cross validation* memiliki maksud bahwa data akan dibagi menjadi 5 bagian sama banyak sehingga masing-masing bagian berjumlah 60 data. Fungsi dari *fold* 5 yaitu untuk menentukan jumlah data testing yaitu:

$$\frac{banyak\ data}{k\ fold} = \frac{300}{5} = 60$$

Sehingga 60 data digunakan sebagai data testing dan jumlah data sisanya digunakan sebagai data training.

4. 10-*fold cross validation*

10-*fold cross validation* memiliki maksud bahwa data akan dibagi menjadi 10 bagian sama banyak sehingga masing-masing bagian berjumlah 30 data. Fungsi dari fold 10 yaitu untuk menentukan jumlah data testing yaitu:

$$\frac{banyak\ data}{k\ fold} = \frac{300}{10} = 30$$

Sehingga 30 data digunakan sebagai data testing dan jumlah data sisanya digunakan sebagai data training.

2.7. Implementasi Naïve Bayes Classification

Naïve bayes classifier yaitu jenis model klasifikasi yang terkenal dan digunakan untuk melakukan penambangan data dikarenakan pengimplementasiannya metode ini tidak rumit [12]. Metode *naïve bayes classifier* ini membutuhkan waktu pemrosesan yang tidak lama [16], dan tidak rumit untuk diterapkan dengan struktur yang cukup sederhana serta struktur efektifitas yang tinggi. *Naïve bayes classifier* ialah suatu metode yang memiliki anggapan yakni ada atau tidak adanya fitur di kelas tidak ada hubungannya dengan keberadaan fitur lain. Meskipun fitur ini tergantung satu fitur dengan fitur lain, namun pengklasifikasikan naïve bayes akan berlanjut untuk mengasumsikan jika fitur-fitur ini independen dan tidak berpengaruh satu sama lain.

$$P(A|B) = \frac{P(A|B).P(A)}{P(B)} \quad (5)$$

Metode klasifikasi *Naive Bayes* masing-masing tweet direpresentasikan melalui pasangan atribut (a1, a2, a3, ..., an) dimana kata pertama berupa a1, kata kedua berupa a2 dan begitu seterusnya, namun himpunan kelas berupa V. Ketika proses klasifikasi, metode ini akan menampilkan hasil kategori/kelas dengan probabilitas tertinggi (VMAP) dengan menginputkan semua atribut (a1, a2, a3, ..., an). Persamaan VMAP dapat dinotasikan sebagai berikut ini:

$$VMAP = \underset{vj \in v}{\arg max} P(v_j|a_1a_2a_3\ldots a_n) \quad (6)$$

Menurut penggunaan teorema bayes, persamaan diatas bisa dinotasikan melalui persamaan berikut ini:

$$VMAP = \underset{vj \in v}{\arg max} \frac{P(a1a2a3\ldots vj)P(vj)}{P(vj|a1a2a3\ldots an)} \quad (7)$$

Nilai P (a1, a2, a3, ..., an) konstan untuk semua vj sehingga persamaan diatas juga bisa dinyatakan sebagai persamaan di bawah ini:

$$VMAP = \underset{vj \in v}{\arg max} P(v_j|a_1a_2a_3\ldots a_n|\ v_j) \quad (8)$$

Nave Bayes Classifier menyederhanakan ini dengan menganggap pada setiap kategori, masing-masing atribut bebas bersyarat satu sama lain. Maka persamaannya:

$$P(v_j|a_1a_2a_3\ldots a_n|\ v_j) = \prod_i P(a_i|v_j) \quad (9)$$

3. Hasil dan Pembahasan

Berikut ini merupakan nilai akurasi yang dihasilkan pada cross validation untuk masing-masing ekstraksi fitur. Teknik klasifikasi yang digunakan dalam semua pengujian menggunakan algoritma Naïve Bayes. Tujuan dari pengujian adalah untuk membandingkan kinerja ekstraksi fitur TF-IDF dan Word2Vec pada berbagai skenario k-fold dan dataset dilihat dari parameter akurasinya.

a. TF-IDF

Nilai akurasi yang dihasilkan oleh ekstraksi fitur TF-IDF yang didapatkan pada proses *cross validation* dapat dilihat pada Tabel 1.

https://doi.org/10.31849/digitalzone.v13i1.9158

Tabel 1. Nilai Akurasi Ekstraksi Fitur TF-IDF

K-fold cross	**Langkah Uji**	**TF-IDF**
2-fold	Langkah Uji 1	0,82
	Langkah Uji 2	0,83
3-fold	Langkah Uji 1	0,86
	Langkah Uji 2	0,75
	Langkah UJi 3	0,87
5-fold	Langkah Uji 1	0,85
	Langkah Uji 2	0,86
	Langkah Uji 3	0,7
	Langkah Uji 4	0,83
	Langkah Uji 5	0,88
10-fold	Langkah Uji 1	0,76
	Langkah Uji 2	0,9
	Langkah Uji 3	0,9
	Langkah Uji 4	0,86
	Langkah Uji 5	0,73
	Langkah Uji 6	0,7
	Langkah Uji 7	0,9
	Langkah Uji 8	0,76
	Langkah Uji 9	0,96
	Langkah Uji 10	0,9

b. Word2Vec

Nilai akurasi yang dihasilkan oleh ekstraksi fitur *wor2vec* yang didapatkan pada proses *cross validation* dapat dilihat pada tabel 2.

Tabel 2. Nilai Akurasi Ekstraksi Fitur *Word2Vec*

K-fold cross	**Langkah Uji**	***Word2Vec***
2-fold	Langkah Uji 1	0,513
	Langkah Uji 2	0,533
3-fold	Langkah Uji 1	0,53
	Langkah Uji 2	0,52
	Langkah UJi 3	0,52
5-fold	Langkah Uji 1	0,566
	Langkah Uji 2	0,483
	Langkah Uji 3	0,516
	Langkah Uji 4	0,466
	Langkah Uji 5	0,583
10-fold	Langkah Uji 1	0,6
	Langkah Uji 2	0,5
	Langkah Uji 3	0,466
	Langkah Uji 4	0,5
	Langkah Uji 5	0,466
	Langkah Uji 6	0,566
	Langkah Uji 7	0,5
	Langkah Uji 8	0,433
	Langkah Uji 9	0,6
	Langkah Uji 10	0,566

Selanjutnya yaitu pengujian data dilakukan dengan cara menguji masing-masing langkah uji pada setiap fold dengan menggunakan 50 *unseen data test* yang sudah ditentukan komposisinya yaitu 25 positif dan 25 negatif. Tabel di bawah ini menampilkan hasil pengujian data untuk setiap ekstraksi fitur, yakni TF-IDF dan *Word2Vec*

Tabel. 3 Pengujian *Unseen Data Test* TF-IDF

K-fold cross	Langkah Uji	TF-IDF	Pengujian Data
2-fold	Langkah Uji 1	0,82	0,84
	Langkah Uji 2	0,83	0,88
3-fold	Langkah Uji 1	0,86	0,86
	Langkah Uji 2	0,75	0,84
	Langkah UJi 3	0,87	0,90
5-fold	Langkah Uji 1	0,85	0,92
	Langkah Uji 2	0,86	0,92
	Langkah Uji 3	0,7	0,88
	Langkah Uji 4	0,83	0,88
	Langkah Uji 5	0,88	0,92
10-fold	Langkah Uji 1	0,76	0,92
	Langkah Uji 2	0,9	0,92
	Langkah Uji 3	0,9	0,90
	Langkah Uji 4	0,86	0,92
	Langkah Uji 5	0,73	0,90
	Langkah Uji 6	0,7	0,92
	Langkah Uji 7	0,9	0,90
	Langkah Uji 8	0,76	0,90
	Langkah Uji 9	0,96	0,92
	Langkah Uji 10	0,9	0,92

Tabel 4. Pengujian *Unseen Data Test Word2vec*

K-fold cross	Langkah Uji	*Word2Vec*	Pengujian data
2-fold	Langkah Uji 1	0,513	0,44
	Langkah Uji 2	0,533	0,44
3-fold	Langkah Uji 1	0,53	0,44
	Langkah Uji 2	0,52	0,44
	Langkah UJi 3	0,52	0,44
5-fold	Langkah Uji 1	0,566	0,44
	Langkah Uji 2	0,483	0,44
	Langkah Uji 3	0,516	0,44
	Langkah Uji 4	0,466	0,44
	Langkah Uji 5	0,583	0,44
10-fold	Langkah Uji 1	0,6	0,44
	Langkah Uji 2	0,5	0,44
	Langkah Uji 3	0,466	0,44
	Langkah Uji 4	0,5	0,44
	Langkah Uji 5	0,466	0,44
	Langkah Uji 6	0,566	0,44
	Langkah Uji 7	0,5	0,44
	Langkah Uji 8	0,433	0,44
	Langkah Uji 9	0,6	0,44
	Langkah Uji 10	0,566	0,44

https://doi.org/10.31849/digitalzone.v13i1.9158

Berdasarkan skenario pengujian yang telah dilakukan terhadap sentiment analisis pada data twitter *review* pelayanan indihome berbasis naïve bayes dengan menggunakan fitur TF-IDF dan juga *word2vec* diperoleh rekapitulasi hasil akurasi yang ditampilkan pada tabel 5 berikut.

Tabel 5. Hasil Perhitungan K-fold cross validation TF-IDF dan Word2Vec

K-fold cross	Langkah Uji	TF-IDF	Pengujian Data	*Word2Vec*	Pengujian data
2-fold	Langkah Uji 1	0,82	0,84	0,513	0,44
	Langkah Uji 2	0,83	0,88	0,533	0,44
3-fold	Langkah Uji 1	0,86	0,86	0,53	0,44
	Langkah Uji 2	0,75	0,84	0,52	0,44
	Langkah UJi 3	0,87	0,90	0,52	0,44
5-fold	Langkah Uji 1	0,85	0,92	0,566	0,44
	Langkah Uji 2	0,86	0,92	0,483	0,44
	Langkah Uji 3	0,7	0,88	0,516	0,44
	Langkah Uji 4	0,83	0,88	0,466	0,44
	Langkah Uji 5	0,88	0,92	0,583	0,44
10-fold	Langkah Uji 1	0,76	0,92	0,6	0,44
	Langkah Uji 2	0,9	0,92	0,5	0,44
	Langkah Uji 3	0,9	0,90	0,466	0,44
	Langkah Uji 4	0,86	0,92	0,5	0,44
	Langkah Uji 5	0,73	0,90	0,466	0,44
	Langkah Uji 6	0,7	0,92	0,566	0,44
	Langkah Uji 7	0,9	0,90	0,5	0,44
	Langkah Uji 8	0,76	0,90	0,433	0,44
	Langkah Uji 9	0,96	0,92	0,6	0,44
	Langkah Uji 10	0,9	0,92	0,566	0,44

Tabel rekapitulasi diatas dapat dijelaskan bahwa perolehan nilai akurasi tertinggi fitur tf-idf yaitu 96% dengan k=10 pada langkah uji kesembilan, sedangkan untuk nilai akurasi tertinggi yang dihasilkan oleh fitur *word2vec* yaitu 60% dengan k=10 pada langkah uji pertama dan kesembilan. Hal ini menunjukkan bahwa tingkat akurasi yang dihasilkan oleh fitur pembobotan kata tf-idf lebih besar daripada tingkat akurasi yang dihasilkan oleh fitur *word2vec*.

4. Kesimpulan

Tujuan penelitian ini adalah untuk membandingkan teknik ekstraksi fitur TF-IDF dan Word2Vec pada data review pelayanan indihome yang diekstraksi dengan algoritma Naïve Bayes. Hasil eksperimen dan validasi dengan teknik k-fold cross validation mengkonfirmasi bahwa TF-IDF lebih baik dengan nilai akurasi mencapai 96% dibandingkan Word2Vec yang mencapai nilai akurasi 60%. Sehingga pada penelitian ini ekstraksi fitur TF-IDF pada penelitian ini mampu menghasilkan nilai akurasi yang lebih besar dibandingkan dengan fitur *Word2Vec*. Saran untuk penelitian selanjutnya supaya nilai akurasi yang dihasilkan oleh fitur *word2vec* bisa bertambah, dapat menggunakan data train dengan jumlah yang besar.

Daftar Pustaka

[1] B. S. Rintyarna, H. Kuswanto, R. Sarno, and E. K. Rachmaningsih, "Modelling Service Quality of Internet Service Providers during COVID-19 : The Customer Perspective Based on Twitter Dataset," pp. 1–12, 2022.

[2] B. S. Rintyarna, "Mapping acceptance of Indonesian organic food consumption under Covid-19 pandemic using Sentiment Analysis of Twitter dataset," J. Theor. Appl. Inf. Technol., vol. 99, no. 5, pp. 1009–1019, 2021.

[3] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Semantic Features for Optimizing Supervised Approach of Sentiment Analysis on Product Reviews," MDPI Comput., vol. 8, no. 3, pp.

1–16, 2019.
[4] F. Rahutomo, D. S. E. Ikawati, and O. A. Rohman, "Evaluasi Fitur Word2Vec Pada Sistem Ujian Esai Online," JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform., vol. 4, no. 1, pp. 36–45, 2019.
[5] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," INTEGER J. Inf. Technol., vol. 1, no. 1, pp. 32–41, 2017, [Online]. Available: https://www.researchgate.net/profile/Ghulam_Buntoro/publication/316617194_Analisis_Sentimen_Calon_Gubernur_DKI_Jakarta_2017_Di_Twitter/links/5907eee44585152d2e9ff992/Analisis-Sentimen-Calon-Gubernur-DKI-Jakarta-2017-Di-Twitter.pdf.
[6] H. Murfi, F. L. Siagian, and Y. Satria, "Topic features for machine learning-based sentiment analysis in Indonesian tweets," Int. J. Intell. Comput. Cybern., p. IJICC-04-2018-0057, 2019, doi: 10.1108/IJICC-04-2018-0057.
[7] A. K. M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, "Sentiment Strength Detection in Short Informal Text," J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 12, pp. 2544–2558, 2010.
[8] F. W. KURNIAWAN, "Analisis Sentimen Twitter Bahasa Indonesia dengan Word2Vec," Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 2, pp. 4704–4713, 2020, [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/159923/slug/analisis-sentimen-twitter-bahasa-indonesia-dengan-word2vec.html%0A/home/catalog/id/159923/slug/analisis-sentimen-twitter-bahasa-indonesia-dengan-word2vec.html.
[9] S. Fransiska and A. Irham Gufroni, "Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method," Sci. J. Informatics, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji.
[10] B. S. Rintyarna, "Pengaruh Seleksi Fitur Pada Skema Klasifikasi Naive Bayes Berbasis Gaussian dan Kernel Density," J. Sist. dan Teknol. Inf. Indones., vol. 1, no. 1, pp. 26–30, 2016, [Online]. Available: file:///C:/Users/User/Downloads/fvm939e.pdf.
[11] S. Choirunnisa, "Metode Hibrida Oversampling Dan Ketidakseimbangan Data Kegagalan," 2019.
[12] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," PLoS One, vol. 14, no. 8, pp. 1–20, 2019, doi: 10.1371/journal.pone.0220976.
[13] M. A. Fauzi, F. Nur, and T. Afirianto, "Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion," TELKOMNIKA (Telecommunication, Comput. Electron. Control., vol. 16, no. 3, pp. 1345–1350, 2018, doi: 10.12928/TELKOMNIKA.v16i3.7751.
[14] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," 2018, [Online]. Available: http://arxiv.org/abs/1811.12808.
[15] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," Comput. Stat., vol. 36, no. 3, pp. 2009–2031, 2021, doi: 10.1007/s00180-020-00999-9.
[16] H. Chen and D. Fu, "An Improved Naive Bayes Classifier for Large Scale Text," vol. 146, no. Icaita, pp. 33–36, 2018, doi: 10.2991/icaita-18.2018.9.

https://doi.org/10.31849/digitalzone.v13i1.9158