IDENTIFIKASI PARAFRASA BAHASA INDONESIA MENGGUNAKAN NAÏVE BAYES

Bayu Indrawarman Julianto[1] , Adiwijaya [3], Mohamad Syahrul Mubarok[3]
#School of Computing, Telkom University
Jl. Telekomunikasi No.01, Terusan Buah Batu, Bandung, Jawa Barat, Indonesia
[1]bayu_i93@yahoo.com
[2]adiwijaya@telkomuniversity.ac.id
[3]msyahrulmubarok@telkomuniversity.ac.id

Abstrak

Salah satu tujuan dari *Natural Languange Processing* adalah mengidentifikasi parafrasa, yang berarti untuk mengajarkan kepada mesin apakah sebuah kalimat memiliki makna yang sama dengan kalimat lainnya. Parafrasa berarti pengungkapan kembali suatu tuturan dari sebuah tingkatan atau macam bahasa menjadi yang lain tanpa merubah pengertian. Dalam penelitian ini dilakukan klasifikasi untuk menentukan apakah dua kalimat Bahasa Indonesia termasuk kedalam parafrasa atau non-parafrasa.. Penelitian dilakukan dengan menggunakan Naïve Bayes sebagai *classifer*. Performansi terbaik dari sistem menghasilkan akurasi 0.713, presisi 0.688, *recall* 0.798, dan *F1-Measure* 0.735.

Kata kunci :*Naive Bayes*, identifikasi parafrasa, *preprocessing.*

Abstract

Paraphrase identification is one of the process in Natural Languange Processing, in purpose to teach the machine if one query is having same meaning to another. The meaning of paraphrase itself is restatement of text or passage using different words. In this research we focusing on classifying two Indonesian sentence whether it is a paraphrase or not. The method for classifier is Naive Bayes. The result from classifier testing as follows: Accuracy 0.713, Precision 0.688, Recall 0.798, and F1-Measure 0.735.

Keyword : *Text Mining, Naive Bayes, Paraphrase Identification.*

1. Pendahuluan

Natural Languange Processing (NLP) adalah bagian dari ilmu komputer, Artificial Intelligence dan kebahasaan yang membahas tentang teknik yang digunakan untuk mengekstraksi struktur kalimat dan arti dari input yang bertujuan untuk membantu mesin melakukan tugas yang diberikan dengan bahasa alam. Ilmu dari NLP sangat luas dikarenakan banyak sekali bahasa yang ada [1].

Contoh dari pemanfaatan NLP dapat ditemukan pada information retrieval, text summarization, question answering, machine translation, serta deteksi plagiarisme. Dalam deteksi plagiarisme, salah satu prosesnya merupakan identifikasi parafrasa. Parafrasa berarti pengungkapan kembali suatu tuturan dan sebuah tingkatan atau macam bahasa menjadi yang lain tanpa merubah pengertian; Parafrasa juga dapat diartikan sebagai penguraian kembali suatu teks dalam bentuk yang lain, dengan maksud untuk dapat menjelaskan makna yang tersembunyi [2]. Parafrasa digunakan sebagai teknik untuk menjelaskan sesuatu menggunakan kalimat yang berbeda namun memiliki makna yang sama.

Untuk mengenali parafrasa, mesin perlu untuk mengenali frasa-frasa yang berbeda namun memiliki makna yang sama. Seperti contoh, dalam kalimat "Bekerja dari pagi" mesin dapat mengenali bahwa kalimat tersebut sejenis dengan "Berbakti dari subuh". Dalam bahasa Indonesia, terdapat prefiks, sufiks, infiks, dan konfiks pada struktur bahasa sehingga sulit untuk menyocokkan kata yang berkaitan.

Untuk menghadapi permasalahan diatas maka dibutuhkan sebuah proses yang dinamakan identifikasi parafrasa. Identifikasi parafrasa adalah proses untuk mengenali apakah sepasang kalimat memiliki makna yang sama atau tidak. Dalam identifikasi parafrasa, dilakukan preprocessing untuk meningkatkan kualitas data. Preprocessing terdiri dari tokenization, normalization, stopword removal dan stemming. Algoritma stemming yang digunakan untuk preprocessing dataset parafrasa bahasa Indonesia adalah algoritma Nazief-Adriani yang memiliki nilai performansi terbaik untuk dataset Bahasa Indonesia [5]. Data hasil preprocessing kemudian dilakukan proses ekstraksi fitur yang bertujuan untuk membangun fitur-fitur baru dari dataset tersebut. Fitur yang pertama adalah fitur sintaktik yang merupakan hasil dari perhitungan jarak antara dua kalimat, perhitungan tersebut menggunakan metode Levhenstein Distance. Fitur kedua adalah fitur semantik, dimana fitur ini menghitung tingkat kemiripan berdasarkan pohon semantik, perhitungan tersebut menggunakan metode Wu and Palmer. Setelah fitur diekstraksi, dataset dibagi menjadi dua yaitu Training set dan testing set. Setelah dibagi, dilakukan diskritisasi data dengan menggunakan metode K-Means. Metode yang digunakan untuk melatih classifier adalah Naïve Bayes.

Naïve Bayes merupakan suatu metode pengklasifikasian dengan probabilitas dan statistik berdasarkan teorema Bayes. Alasan penggunaan metode Naïve Bayes yaitu tidak memperlukan data training yang banyak, dapat menangani data yang hilang, dan dapat digunakan dengan fitur yang independen [4].

2. Tinjauan Pusaka

2.1 Parafrase

Parafrasa adalah penguraian kembali suatu teks dalam bentuk yang lain, dengan maksud untuk dapat menjelaskan makna yang tersembunyi [2]. Dalam parafrase, dibutuhkan aturan yang ketat, untuk membuktikan apakah kalimat tersebut merupakan sebuah parafrase atau bukan.

2.2 Teorema Bayes

Teorema Bayes merupakan teorema yang digunakan dalam statistika untuk menghitung peluang untuk suatu kejadian, berdasarkan kondisi yang mungkin mempengaruhi kejadian tersebut [6].

Secara matematis, teorema bayes adalah:

$$P(A|B) = \frac{P(B|A)\times P(A)}{P(B)} \quad (1)$$

Dimana:
A dan B merupakan kejadian
P(A) dan P(B) merupakan *Prior Probabilities*, atau probabilitas terjadinya A dan B tanpa ada pengaruh dari hal lain
P(B|A) merupakan *Conditional Probabilities*, atau probabilitas terjadinya kejadian B dipengaruhi kejadian A
P(A|B) merupakan *Posterior Probabilities*, atau probabilitas terjadinya A, yang dipengaruhi oleh *Conditional Probabilities* dari B terhadap A.

2.3 *Multinomial Naïve Bayes*

Multinomial Naïve Bayes (MLB) adalah metode *Bayesian Learning* yang dikembangkan dari metode *Naïve Bayes*. Metode ini berasal dari *teorema Bayes*, dan menggunakan klasifier statistik berdasarkan peluang. Teorema ini menggunakan distribusi multinomial pada fungsi *Conditional Probabilities.*

Secara perumusan, perhintungan *posterior probability* dari MLB dijabarkan sebagai:

$$p(C_k|x_1,..,x_n) = \frac{p(C_k)p(x_1,..,x_n|C_k)}{p(x_1,..,x_n)} \quad (2)$$

Disini, asumsi "naïve" dari *Naïve Bayes*, yaitu asumsi independent dimana setiap fitur x merupakan independen dari fitur lainnya, berlaku. Hal itu menyebabkan:

$$\begin{aligned} p(C_k|x_1,..,x_n) &\propto p(C_k)p(x_1,..,x_n|C_k) \\ &\propto p(C_k)p(x_1|C_k)p(x_2|C_k)p(x_3|C_k) \ldots . p(x_n|C_k) \\ &\propto p(C_k)\prod_{i=1}^{n} p(x_i C_k) \end{aligned} \quad (3)$$

Oleh karena itu, untuk menghitung nilai *Posterior Probability* terbesar menjadi :

$$\hat{y} = \arg\max_{k\in\{1,..,K\}} p(C_k)\prod_{i=1}^{n} p(x_i C_k) \quad (4)$$

Dimana *x* merupakan model fitur dengan *n* merepresentasikan setiap fiturnya, untuk setiap *k* kemungkinan dari kelas C_k.

3. Metode dan Perancangan Sistem

3.1 Dataset

Data yang digunakan dalam penelitian ini menggunakan data teks berbahasa Indonesia yang berupa pasangan kalimat / frasa. Total dataset berjumlah 1004 data.

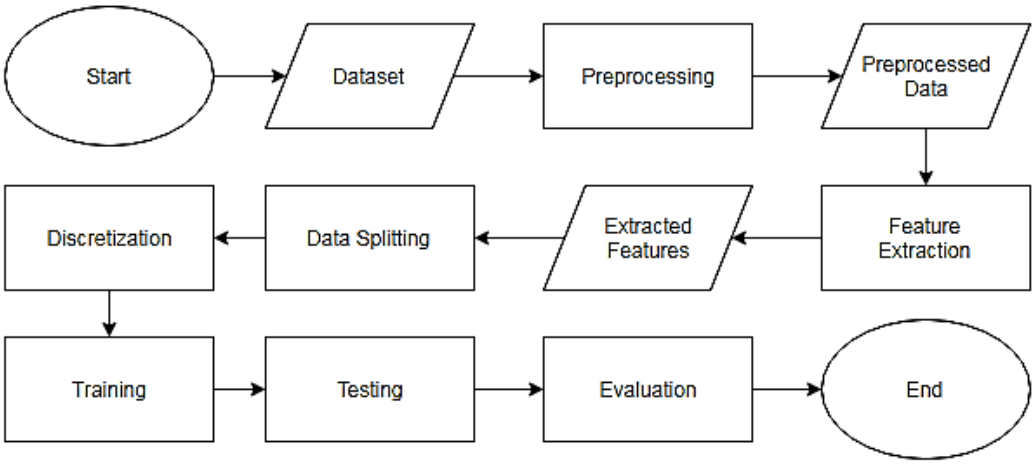
Dataset terdiri dari 3 kolom dalam setiap barisnya, dimana kolom pertama berisi kalimat 1, kolom kedua berisi kalimat 2, dan kolom ketiga berisi kelas yang terdiri dari dua kelas yaitu 0 dimana kedua kalimat tersebut merupakan non parafrasa dan 1 yang berarti kedua kalimat tersebut merupakan parafrasa. Tabel 1 menggambarkan distribusi dataset.

Tabel 1. Distribusi Dataset

Kelas	Jumlah Data	Persentase
Parafrasa	502	50%
Non Parafrasa	502	50%
Total Dataset	1004	100%

3.2 Alur sistem keseluruhan

Gambaran umum sistem identifikasi parafrasa Naive Bayes digambarkan dalam gambar 1.



Gambar 1. Rancangan Umum Sistem

Secara umum, terdapat proses yang berperan penting dalam penelitian ini, yaitu *Preprocessing*, *Feature Extraction*, *Data Splitting*, *Discretization*, Proses *Training* Naive Bayes dan *Testing* menggunakan *classifier* yang telah dilatih sebelumnya.

3.3 Preprocessing

Preprocessing yang digunakan dalam penelitian ini berupa *tokenization*, normalisasi, *stop word removal*, dan *stemming* dengan menggunakan metode Nazief-Andriani yang berupa metode *stemming* khusus bahasa Indonesia.

3.4 Feature Extraction

Ekstraksi fitur dilakukan dengan dua jenis fitur, yang pertama adalah fitur sintaktik yang merupakan hasil dari perhitungan jarak antara dua kalimat, dan diukur menggunakan metode *Levhensthein Distance*. Fitur yang kedua adalah fitur semantik, yaitu fitur yang didapatkan dari jarak kemiripan antara dua kalimat berdasarkan pohon semantik. Fitur semantik diukur menggunakan metode *Wu and Palmer*.

Setelah fitur ekstraksi didapatkan, kemudian dilakukan cosine similarity terhadap vektor untuk masing-masing fitur. Cosine Similarity digunakan untuk menemukan kesamaan antara dua kalimat.

3.5 Data Splitting

Dataset yang sudah melewati proses ekstraksi fitur kemudian dibagi kedalam dua bagian yaitu *Training set* dan *Testing set*. Untuk membagi data yang ada, penulis membagi kedalam tiga skenario porsi pembagian dataset yang akan dijelaskan dalam analisis hasil.

3.6 Discretization

Diskritisasi adalah prosedur yang dapat memproses data yang merubah data kualitaitif menjadi data kuantitatif. Dalam penelitian ini, diskritisasi dilakukan dengan menggunakan metode *K-Means Clustering.* Terdapat enam skenario nilai k untuk pengujian, dengan masing-masing nilai k yaitu 2,5,8,11,14, dan 17.

3.7 Training Classifier.

Dari dataset yang telah diolah, maka *classifier Naive Bayes* dapat dibangun. Dalam proses ini, dilakukan *training* dengan data *training* yang telah dibagi sebelumnya.

3.8 Testing and Evaluation

Classifier yang dibangun kemudian diujikan dengan parameter berupa nilai *Accuracy, Precision, Recall,* dan *F1-Measure.*

4. Analisis dan Hasil Pengujian

4.1 Tujuan Pengujian

Tujuan dari pengujian yang dilakukan sebagai berikut.

a. Menganalisis pengaruh dari persentase pembagian data *training* dan data *testing* terhadap hasil identifikasi parafrasa *Naïve Bayes* dengan menggunakan nilai kontinu.
b. Menganalisis pengaruh dari nilai k dalam diskritisasi dengan menggunakan *K-Means Clustering* terhadap hasil identifikasi parafrasa *Naïve Bayes.*

4.2 Skenario Pengujian

Skenario Pengujian yang dilakukan adalah sebagai berikut.

a. Rasio pembagian dataset
Dataset yang ada dibagi kedalam 3 skenario rasio yaitu 20% data *Training* - 80% data *Testing*, 50% data *Training* - 50% data *Testing*, dan 80% data *Training* - 20% data *Testing*, dimana data *Testing* dari ketiga skenario adalah data yang sama sebanyak 200 data dan jumlah data *Training* berubah sesuai dengan skenarionya yaitu 50 data untuk rasio 20%, 200 data untuk rasio 50%, dan 800 data untuk rasio 80%. Dengan catatan nilai yang digunakan dalam skenario ini adalah nilai kontinu / tidak dilakukan diskritisasi.
b. Nilai *k* dalam diskritisasi dengan *K-Means Clustering*
Pengujian dilakukan dengan menggunakan enam nilai k yaitu 2, 5, 8, 11, 14, dan 17. Pengujian juga dilakukan dengan membandingkan rasio pembagian dataset seperti yang dijelaskan di skenario pembagian dataset.
Seluruh skenario diujikan dengan data *training* diambil secara acak dan diulang sebanyak masing-masing 5 kali untuk melihat nilai rata-rata dari pengujian.

4.3 Hasil Pengujian dan Analisis

a. Analisis rasio pembagian dataset
Berikut adalah tabel hasil nilai rata-rata pengujian pengaruh rasio pembagian dataset.

Tabel 2. Analisis rasio pembagian dataset

Evaluation	**20:80**	**50:50**	**80:20**
Accuracy	0.707	0.720	0.742
Precision	0.803	0.797	0.803
Recall	0.576	0.602	0.634
F1-Measure	0.655	0.675	0.710

Berdasarkan tabel 2 , dilihat bahwa distribusi 80:20 memiliki nilai performansi terbaik. Hal ini disebabkan model memiliki perbandingan data training yang lebih banyak untuk membangun classifier. Naïve bayes merupakan supervised machine learning, hal tersebut berarti banyaknya data untuk membangun classifier berpengaruh besar terhadap kinerja dari model untuk melakukan klasifikasi. Semakin banyak data training yang dimiliki maka kemampuan mesin untuk mempelajari pola semakin tinggi.

b. Analisis nilai *k* dalam diskritisasi
Berikut adalah tabel hasil nilai rata-rata pengujian pengaruh nilai k dalam diskritisasi.

Tabel 3. Analisis nilai *k* dalam diskritisasi.

Evaluation	k=2	k=5	k=8	k=11	k=14	k=17
Accuracy	0.713	0.731	0.720	0.694	0.695	0.687
Precision	0.688	0.785	0.755	0.721	0.724	0.692
Recall	0.798	0.644	0.658	0.643	0.628	0.647
F1	0.735	0.704	0.701	0.677	0.675	0.659

Berdasarkan tabel 3, dilihat berdasarkan nilai dari 6 buah k yang berbeda yaitu 2, 5, 8, 11, 14, dan 17 bahwa performansi F1-Measure rata-rata terbesar dimiliki oleh nilai k = 2. Hal ini disebabkan semakin banyaknya nilai k membuat data yang seharusnya dimiliki oleh centroid tertentu malah diambil oleh centroid lainnya. Hal tersebut dapat mengakibatkan terjadinya kesalahan klasifikasi, dimana yang seharusnya merupakan kelas tertentu salah diklasifikasikan kedalam kelas yang lain.

5 Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, maka kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut.

1. Metode klasifikasi Naïve Bayes teruji dapat melakukan identifikasi parafrasa Bahasa Indonesia dengan nilai performansi yaitu akurasi 0.713, presisi 0.688, recall 0.798, dan F1-Measure 0.735.

2. Proses preprocessing dataset Bahasa Indonesia dapat dilakukan, dengan proses stemming menggunakan metode Nazief-Adriani.
3. Proses ekstraksi fitur dilakukan dengan menggunakan Levhenstein Distance untuk fitur sintaktik dan Wu and Palmer untuk fitur sematik.
4. Distribusi 80:20 merupakan distribusi dengan nilai terbaik untuk Naïve Bayes.
5. Nilai k terbaik untuk proses diskritisasi adalah k = 2.

5.2 Saran

Sistem yang telah dibangun masih dapat dikembangkan lebih lanjut. Hal yang disarankan diantaranya memperbaiki proses preprocessing yang masih dapat disempurnakan, serta penggunaan dan penambahan metode feature extraction yang lain. Diharapkan dengan ditambahkannya fitur lainnya dapat meningkatkan performansi sistem.

Daftar Pustaka

[1] A. Reshamwala, D. Mishra and P. Pawar, "Review On Natural Language Processing," ACST – Engineering Science and Technology: An International Journal (ESTIJ), vol. 3, no. 1, 2013.
[2] KBBI, “KBBI – Parafrasa” [Online]. Avaliable: kbbi.web.id/parafrasa.
[3] Agusta, L.2009. Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief dan Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. Konferensi Nasional Sistem dan Informatika 2009.
[4] Rish, I (2001), “An empirical study of the Naïve Bayes classifier”.
[5] Asian, Jelita (2009), “Effective Techniques for Indonesian Text Retrieval”
[6] Rachka, Sebastian (2014), “Naïve Bayes and Text Classification I : Introduction and theory”.
[7] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). Tackling the poor assumptions of Naive Bayes classifiers
[8] Cambria, Erik; White, Bebo (2014). "Jumping NLP Curves: A Review of Natural Language Processing Research"
[9] Adiwijaya, 2014, Aplikasi Matriks dan Ruang Vektor, Yogyakarta: Graha Ilmu
[10] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta
[11] MS Mubarok, Adiwijaya, MD Aldhi, 2017, Aspect-based sentiment analysis to review products using Naïve Bayes, AIP Conference Proceedings 1867, 020060 (2017)
[12] Aziz, R.A., Mubarok, M.S. and Adiwijaya, 2016. Klasifikasi topik pada Lirik Lagu dengan Metode Multinomial Naïve Bayes. In Indoensia Symposium on Computing (IndoSC) 2016.
[13] Arifin, A.H.R.Z., Mubarok, M.S. and Adiwijaya, 2016, Learning Struktur Bayesian Netwroks menggunakan Novel Modified Binary Differential Evolution pada Klasifikasi Data. In Indonesia Symposium on Computing (IndoSC) 2016.