Jurnal Politeknik Caltex Riau
Terbit Online pada laman https://jurnal.pcr.ac.id/index.php/jkt/
| **e- ISSN : 2460-5255 (Online) | p- ISSN : 2443-4159 (Print) |**

Implementasi Ekstraksi Fitur untuk Pengelompokan Dokumen Proposal Menggunakan Algoritma Naïve Bayes

Dini Nurmalasari[1], **Heri Ribut Yuliantoro**[2]

[1]Politeknik Caltex Riau, email: dini@pcr.ac.id
[2]Politeknik Caltex Riau, email: heriry@pcr.ac.id

[1] Abstrak

Text mining merupakan proses penemuan informasi yang baru yang belum diketahui sebelumnya dari beberapa dokumen teks. Text mining dapat diterapkan pada bidang ekstraksi informasi, pelacakan topik, perangkuman dokumen, membuat kategorisasi atau pengelompokan dokumen, concept linking atau question answering system. Salah satu yang sering dilakukan dalam mengimplementasikan text mining adalah ekstraksi informasi. Extraksi informasi bertujuan untuk mengambil informasi dari dokumen yang tidak terstruktur menjadi data yang terstruktur, dengan tujuan untuk memudahkan melakukan analisis dari data tersebut. Aktifitas Pengabdian kepada Masyarakat (PkM) diawali dengan pembuatan proposal yang diunggah ke dalam system informasi dalam bentuk dokumen teks. Untuk melihat informasi topik, lokasi dan informasi lainnya dari dokumen tersebut harus dibuka satu per satu, sehingga memakan waktu. Pada penelitian ini ektraksi fitur akan digunakan untuk mengambil fitur dari dokumen proposal PkM dengan menggunakan algoritma Frequent Itemset Mining (FIM). Fitur yang diambil adalah Judul , Abstrak, Tahun Pengabdian, Lokasi, dan topik penelitian. Setelah didapatkan fitur akan dilakukan pengelompokkan topik pengabdian dengan menggunakan algoritma Naive Bayes. Hasil dari penelitian ini dilakukan pengujian dengan confusion matrix, dengan hasil akurasi sebesar 70%. Faktor yang mempengaruhi hasil akurasi diantaranya adalah jumlah data training, sebaran data taining, dan optimasi algoritma yang digunakan. Hasil pengelompokkan ini sangat bermanfaat bagi pengelola aktifitas PkM dalam melihat sebaran data terkait topik, mitra, lokasi dan lain sebagainya.

Kata kunci: *Text Mining, Ektraksi Fitur, Algoritma Naive Bayes*

[2] Abstract

Text mining is the process of discovering new, previously unknown information from several text documents. Text mining can be applied to the fields of information extraction, topic tracking, document summarization, document categorization or grouping, concept linking or question answering systems. One thing that is often done in implementing text mining is information extraction. Information extraction aims to retrieve information from unstructured documents into structured data, with the aim of facilitating analysis of the data. Community Service Activities (PkM) begin with making a proposal that is uploaded to the information system in the form of a text document. To view topic information, location and other information from the document must be opened one by one, so it takes time. In this study, feature extraction will be used to extract features from the PkM proposal document using the Frequent Itemset Mining (FIM) algorithm. The features taken are Title, Abstract, Year of Service, Location, and

Dokumen diterima pada 14 April, 2022
Dipublikasikan pada 26 Juni, 2022

research topic. After obtaining the features, the service topics will be grouped using the Naive Bayes algorithm. The results of this study were tested using a confusion matrix, with an accuracy of 70%. Factors that affect the accuracy results include the amount of training data, the distribution of training data, and the optimization of the algorithm used. The results of this grouping are very useful for PkM activity managers in viewing the distribution of data related to topics, partners, locations etc.

Keywords: *Text Mining, Feature Extraction, Naive Bayes Algorithm*

1. Pendahuluan

Text mining merupakan proses penemuan informasi yang baru yang belum diketahui sebelumnya dari beberapa dokumen teks. Text mining dapat diterapkan pada bidang ekstraksi informasi, pelacakan topik, perangkuman dokumen, membuat kategorisasi atau pengelompokan dokumen, *concept linking* atau *question answering system*. Salah satu yang sering dilakukan dalam mengimplementasikan text mining adalah ekstraksi informasi. Ekstraksi informasi bertujuan untuk mengambil informasi dari dokumen yang tidak terstruktur menjadi data yang terstruktur, dengan tujuan untuk memudahkan melakukan analisis dari data tersebut [1].

Dokumen proposal pelaksanaan pengabdian kepada masyarakat (PkM) merupakan salah satu dokumen tidak terstruktur yang setiap tahunnya terus bertambah di database BP2M Politeknik Caltex Riau, karena setiap Dosen memiliki kewajiban melaksanakan program pengabdian kepada masyarakat sebagai beban kerja dosen setiap semesternya. Sehingga sebagai bentuk implementasi dari kebutuhan tersebut, BP2M setiap tahunnya menyediakan dana dan proses seleksi proposal pengabdian kepada masyarakat. Untuk mempermudah melakukan proses seleksi terhadap dokumen yang masuk, dibutuhkan system yang dapat secara cepat dan tepat membantu untuk mengekstraksi informasi penting dalam dokumen tersebut. Informasi tersebut diantaranya adalah topik pengabdian, mitra kerja sama, biaya yang dibutuhkan, tujuan pengabdian dan lain sebagainya. Sehingga dapat memudahkan dalam melakukan seleksi dan pengelompokan tema pengabdian setiap tahunnya.

Metode yang digunakan untuk melakukan ekstraksi dokumen proposal adalah Text Mining. Text Mining memiliki definisi menambang data berupa teks, bertujuan mencari kata–kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen dengan menggunakan algoritma machine learning, salah satunya adalah algoritma naïve bayes. Algoritma ini dipilih karena dinilai cocok untuk text classification dan berdasarkan kasus serupa, algoritma ini memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma lainnya.

Pada penelitian ini dokumen yang akan digunakan adalah dokumen proposal PkM, dengan tujuan untuk mengambil infomasi penting dari dokumen proposal tersebut secara otomatis untuk kemudian dilakukan pengelompokan, yang nantinya diharapkan dapat membantu dalam proses seleksi proposal. Selain itu penelitian ini juga bertujuan untuk melihat sebaran data topik PkM, sehingga dapat menentukan topik dan mitra berikutnya, sehingga harapannya program PkM ini dapat dimanfaatkan lebih luas.

2. Tinjauan Pustaka

Penelitian mengenai ekstraksi informasi telah banyak dilakukan, baik ektraksi informasi dari dokumen, dari halaman web, maupun dari dokumen gambar dan suara. Penelitian yang pernah dilakukan diantaranya melakukan ekstraksi data yang ditampilkan dalam bentuk list dan tabel yang berasal dari halaman web [2]. *Unsupervised learning algorithms* digunakan dalam penelitian ini untuk memperoleh struktur dari list dan tabel yang ada di web. Penelitian serupa dilakukan oleh [3], melakukan ektraksi data pada table dari halaman web dengan

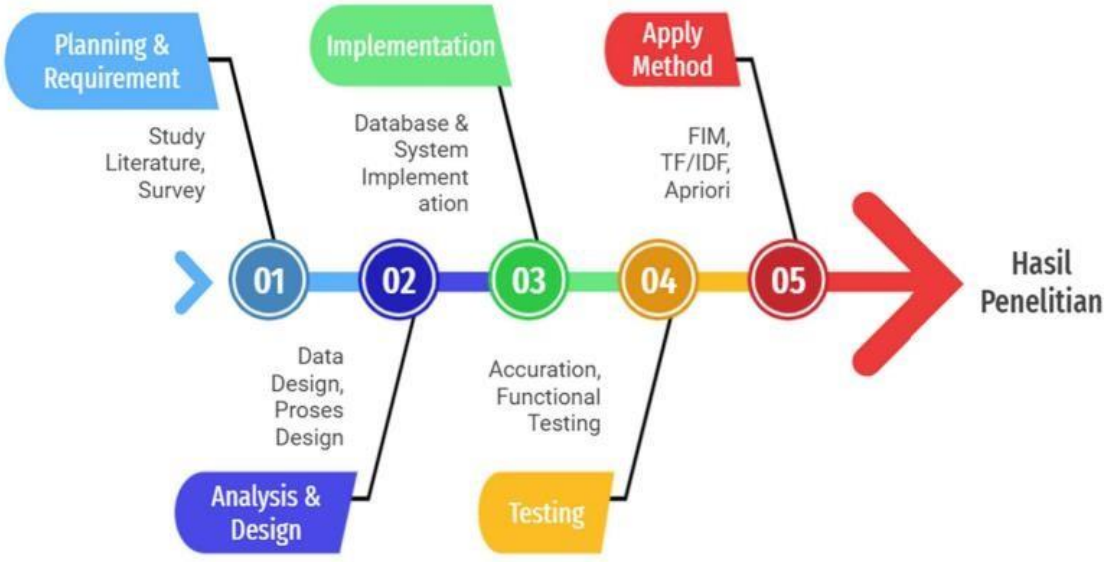
menggunakan pohon DOM. Langkah yang dilakukan adalah dengan mentransformasikan dokumen HTML menjadi pohon DOM kemudian dengan menggunakan DFS (*Depth First Search)* dilakukan penelusuran data.

Penelitian mengenai ekstraksi informasi dari dokumen telah dilakukan oleh [4] [1] [2] dengan membangun *rule-based classification* untuk melakukan ekstraksi informasi dari dokumen Laporan Keuangan. Hal serupa pernah dilakukan oleh [5] dengan menggunakan KNN. Hasilyang diperoleh dari penelitian tersebut yaitu mampu melakukan identifikasi kinerja perusahaan secara otomatis dengan akurasi 85%.

Sedangkan naïve bayes banyak digunakan untuk melakukan klasifikasi teks seperti yang dilakukan oleh [6], menggunakan Naïve Bayes untuk mengelompokan teks berita dan abstrak akademis dengan akurasi yang baik yaitu 91%. Penelitian selanjutnya dilakukan oleh [7], dengan membandingkan metode Naïve Bayes dan *K-Medoids* untuk mengklasifikasikan dokumen tugas akhir dengan tujuan untuk mengelompokkan bidang keahlian. Pada penelitian tersebut didapatkan hasil bahwa Naïve Bayes menghasilkan akurasi lebih baik dibandingan dengan *K-Nearest Neighbored.*

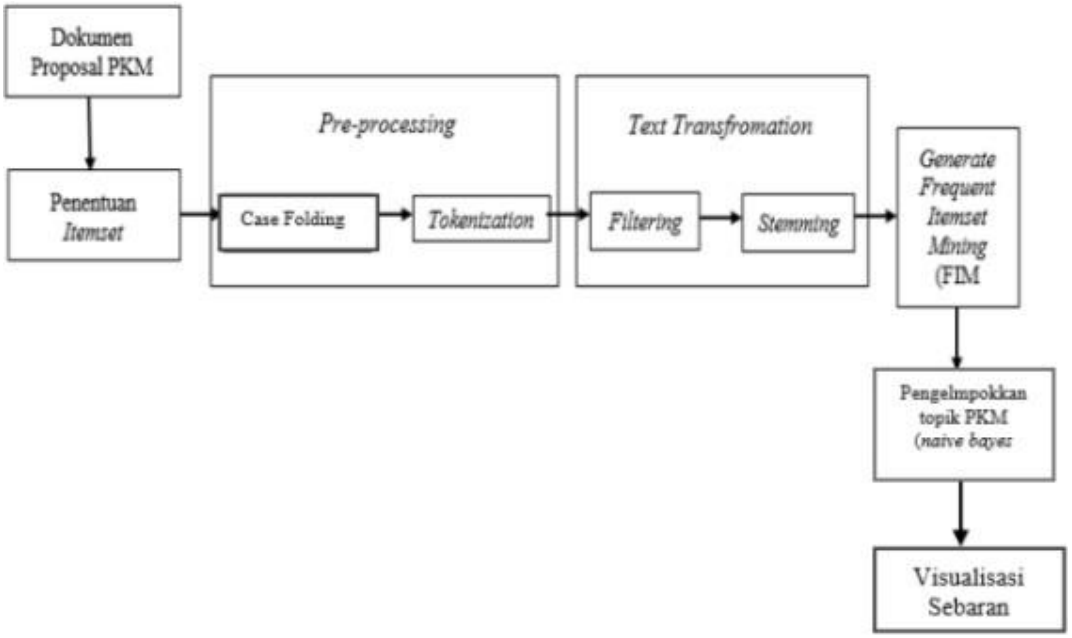
3. Metode Penelitian

Metodologi penelitian yang akan dilakukan mengacu pada tahapan SDLC (*Software Development Life Cycle)*, yang terdiri dari tahapan *requirement* untuk menentukan bisnis proses dan permasalahan data, analysis dan perancangan untuk memetakan hasil dari tahapan sebelumnya kedalam model data yang sesuai, implementasi yaitu tahapan membangun system, serta pengujian.



Gambar 1 Tahapan Penelitian

Sedangkan secara khusus proses pembuatan aplikasi pengelompokan dokumen ini diawali dengan proses upload dokumen Proposal PkM, melakukan penentuan *itemset*, melakukan pembersihan data dengan *case folding*, *tokenizing*, *filtering* dan *stemming*. Selanjutnya ekstraksi fitur dengan men-generate fitur menggunakan FIM (*frequent itemset Mining*). Selanjutnya melakukan pengelompokan topik PkM dengan menggunakan algoritma *Naïve Bayes*, dan terakhir dilakukan visualisasi untuk memudahkan dalam membaca hasil analisis. Pada penelitian ini jenis visualisasi yang akan digunakan adalah *word cloud*, grafik area dan grafik batang. Gambar 2 berikut ini merupakan alur dari proses ekstraksi fitur dari dokumen proposal PkM.



Gambar 2 Implementasi Sistem

4. Hasil dan Pembahasan

Sumber data yang digunakan pada penelitian ini merupakan data proposal PKM Politeknik Caltex Riau dalam format dokumen .doc atau .pdf sebanyak 114 dokumen. Fitur teks yang akan diambil dari dokumen tersebut adalah judul PkM, abstrak, tahun pengabdian, lokasi, dan topik penelitian. Pemilihan fitur tersebut berdasarkan hasil wawancara yang dilakukan dengan kepala Unit Penelitian dan Pengabdian Masyarakat Politeknik Caltex Riau. Berdasarkan tahapan implementasi system pada Gambar 2 langkah yang dilakukan dalam implementasi system ini adalah penentuan *itemset, Preprocessing* dan *text transformation, g*enerate fitur, proses pengelompokan dan visualisasi.

4.1 Preprocessing dan Text Transformation

Pre-processing dan text transformation bertujuan untuk membersihkan data *(data cleaning)* dari dokumen proposal PKM. Proses yang dilakukan adalah *case folding* yaitu merubah semua huruf menjadi huruf kecil dengan tujuan konsistensi, selanjutnya pemotongan kalimat menjadi kata melalui proses *tokenization*, eleminasi kata-kata dengan nilai informasi rendah melalui proses *filtering*, serta pengembalian kata kedalam bentuk dasar melalui proses *stemming*. Tujuan dari proses yang telah disebutkan sebelumnya adalah untuk mengatasi permasalahan *high dimensionality* yaitu besarnya ruang fitur atau dikenal dengan istilah *dimension reduction.* Penguranan dimensi tersebut diharapkan dapat meningkatkan efektifitas dan efisiensi proses *generate feature*.

Pada tahapan *tokenizing* dilakukan proses sebagai berikut [1][10][11]:

1. Mengambil index class pada wordTokenization
2. Menampung sekumpulan ‘word’ dari bagOfWords dan ‘count’ sebagai pfrekuensi pada wordTokenization [index] [tokenizeWords]
3. Memisahkan setiap kata pada data yang sudah di filter berdasarkan class dan ditampung ke dalam splits.

```
(1)     foreach ($this->class as $item) {
(2)     $classData = $this->getDataByClass($item);
(3)     $index = $this->findWordsByClassIndex($item);
(4)     foreach ($this->bagOfWords as $word) {
(5)     $this->wordTokenization[$index]['tokenizeWords'][] = ['word' => $word, 'count' => 0];
(6)        }
(7)     foreach ($classData as $item) {
(8)     $splits = explode(" ", $item['abstract']);
(9)     foreach ($this->wordTokenization[$index]['tokenizeWords'] as $key => $word) {
(10)    foreach ($splits as $split) {
(11)    if ($word['word'] === $split) {
(12)    $this->wordTokenization[$index]['tokenizeWords'][$key]['count']++;
(13)    }
(14)        }
(15)            }
(16)                }
```

Gambar 3 Potongan Kode Program Proses Tokenizing

Tahapan *Filtering* pada penelitian ini digunakan untuk menghapus kata yang dianggap tidak bermakna melalui referensi *stopword*, sementara tahapan *Stemming* digunakan untuk merubah setiap kata kedalam bentuk kata dasar (*stemmed)[13].*

```
(1)     foreach ($this->data as $index => $item) {
(2)     $stopworded = $stopworder->remove($item['abstract']);
(3)     $stemmed = $stemmer->stem($stopworded);
(4)     $this->stemmedData[] = $stemmer->stem($stopworded);
(5)     $this->data[$index]['abstract'] = $stemmed;
(6)       }
(7)     $this->setWords($stemmer->getStemWords());
```

Gambar 4 Potongan Kode Program Proses *Filtering & Stemming*

4.2 Generate Feature

Frequent Itemset Mining merupakan suatu metode yang digunakan untuk melakukan pencarian kata yang *frequent* pada dokumen. Algoritma yang digunakan adalah algoritma Apriori. Pencarian fitur kata menggunakan algoritma Apriori diawali dilakukan dengan Langkah-langkah berikut ini [7][9][14] :

1. Mencari bobot kemunculan kata (term) terhadap dokumen
2. Menghitung kemiripan *vector query* dalam setiap dokumen yang ada
3. Menghitung panjang *vector* dengan mengkuadratkan bobot setiap term dalam setiap dokumen
4. Menerapkan rumus *cosine similarity*
5. Menghitung relevansi antar *query*
   Mengurutkan hasil

4.3 Pengelompokan dengan *Naïve Bayes*

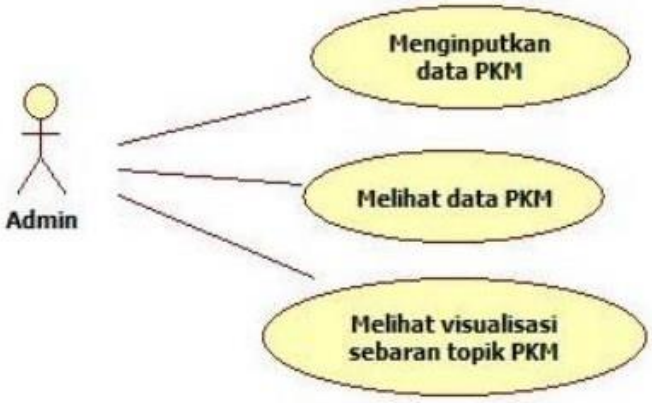
Metode Naïve Bayes diimplementasikan untuk melakukan pengklasifikasian topik dari data yang sudah melalui tahap text mining[17]. Secara umum algoritma Naïve Bayes diawali dengan melakukan perhitungan probabilitas uji kelas berdasarkan hitungan dari setiap kata yang sama[18]. Selanjutnya dilakukan proses pencarian nilai maksimal dari hasil uji kelas kemudian membandingkan nilai maksimal tersebut dengan hasil uji kelas. Berikut implementasi program untuk menentukan klasifikasi topik[16]:

```
(1)     foreach ($this->testClass as $key => $value) {
(2)     foreach ($value['computed'] as $val) {
(3)     $this->testClass[$key]['result'] *= $val;
(4)     }
(5)             }
(6)     $result = [];
(7)     foreach ($this->class as $class) {
(8)     $result[] = $this->testClass[$class]['result'];
(9)       }
(10)    $max = max($result);
(11)    foreach ($this->testClass as $key => $item) {
(12)    if ($item['result'] === $max) return $key;
(13)    }
(14)         return false;
(15)            }
```

Gambar 5 Potongan Kode Program Naïve Nayes

4.4 Visualisasi dalam bentuk Dashboard

Sebelum data hasil analisis divisualisasikan, terdapat beberapa antar muka lainnya yang dibangun pada system ini yaitu menu untuk menginputkan data dokumen PkM atau *load data*, melihat data PkM atau melakukan pencarian data, dan melihat hasil analisis dalam bentul visualisasi dashboard. Fungsionalitas system tersebut digambarkan dalam bentuk *use case diagram* pada gambar 6 berikut ini :

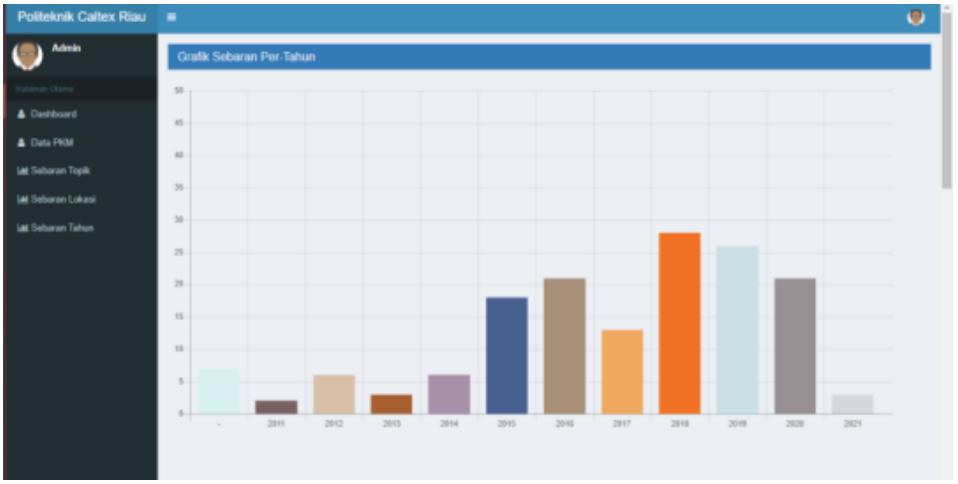


Gambar 6 Use Case Diagram

Visualisasi yang dipilih pada penelitian ini terdiri dari grafik batang, grafik summary text dan *wordcloud.* Visualisasi ini digunakan untuk menampilkan hasil pengelompokan data berupa sebaran topik pengabdian berdasarkan program studi, lokasi, dan waktu. Contoh hasil visualisasi dalam bentuk grafik dapat dilihat pada gambar 7, gambar 8 dan gambar 9 berikut ini.

Gambar 7 Sebaran Data PkM berdasarkan topik

Gambar 8 Visualisasi dengan Word Cloud



Gambar 9 Sebaran data dalam bentuk Grafik Batang

Berdasarkan hasil pengujian yang dilakukan dengan menggunakan *Confusion Matrix* terhadap 114 data training dan 30 data uji, ditemukan hasil pengelompokan yang tidak sesuai, seperti contoh berikut ini :

Table 1 Data Uji

No	Judul PkM	Hasil Klasifikasi (Manual)	Hasil Klasifikasi (Sistem)
1	Alat Resusitasi Jantung Paru	Elektronika	Elektronika
2	Sistem Akses Parkir Dengan QR Code	Elektronika	Komputer
3	Analisis Pembayaran Rekening Air pada UPTD Spam Dinas Pekerjaan Umum dan Penataan Ruang, Perumahan	Komputer	Komputer
4	Rancang Bangun Sistem Informasi Geografis (GIS) Kost / Rumah Sewa Wilayah Kota Stabat Berbasis eb	Komputer	Elektronika
5	Aplikasi Multimedia Mengenal nama-nama Pahlawan Nasional pada SDN -26 Pekanbaru menggunakan Macrome	Komputer	Komputer
6	Robot Penjejak Ruangan Dengan Sensor Ultrasonik dan Kendali Ganda Melalui Bluetooth	Robotika	Robotika

Pada tabel 2 terdapat contoh data dengan hasil pengelompokan yang tidak sesuai antara pengelompokan secara manual berdasarkan pengetahuan dari kepala BP2M dengan hasil prediksi Naïve Bayes. Hasil tersebut didapatkan melalui perhitungan akurasi dan laju error dengan menggunakan *confusion matrix.* Pada tabel 1 berikut ini merupakan hasil perhitungannya dengan menggunakan rumus berikut :

$$Akurasi = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ benar}{Jumlah\ pediksi\ yang\ dilakukan}$$

$$= \frac{21}{30}\ \%$$

$$= 70\ \%$$

$$Laju\ Error = \frac{Jumlah\ data\ yang\ diprediksi\ salah}{Jumlah\ pediksi\ yang\ dilakukan}$$

$$= \frac{9}{30}\ \%$$

$$= 30\%$$

Table 2 Perhitungan *Confusion Matrix*

fij		Data Prediksi						Total Prediksi
		Elektronika	Komputer	Robotika	Umum	Mekatronika	Akuntansi	
Data Aktual	Elektronika	3	1	0	0	0	0	4
	Komputer	1	6	0	0	0	0	7
	Robotika	0	2	3	0	0	0	5
	Umum	0	2	0	3	0	0	5
	Mekatronika	0	1	0	0	3	0	4
	Akuntansi	0	1	0	1	0	3	5
Total Aktual		4	13	3	4	3	3	30

Berdasarkan pengujian yang telah dilakukan, hasil pengelompokan data dokumen proposal PkM dengan menggunakan algoritma Naïve Bayes menghasilkan akurasi sebesar 70%, dengan laju kesalahan sebesar 30%. Setelah dilakukan analisis terhadap faktor yang menyebabkan nilai akurasi tersebut adalah jumlah dataset, serta representasi data yang digunakan. Data training yang digunakan belum bervariasi, sebagai contoh penggunaan kata 'Sistem' pada judul PkM masih mengacu pada kelompok 'Komputer'.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Kesimpulan dari penelitian ini adalah :

1. Akurasi pengelompokan dengan menggunakan perhitungan *confusion matrix* sebesar 70%, dengan laju kesalahan sebesar 30%.
2. Hasil akurasi dipengaruhi oleh data training yang yang masih sedikit, kurang meratanya sebaran data training, optimasi proses ektraksi fitur serta implementasi algoritma pengelompokkan

5.2 Saran

Saran untuk penelitian ini adalah meningkatkan akurasi pada sistem ini dengan memperbaiki dan meningkatkan faktor-faktor yang mempengaruhi akurasi, serta menambahkan fitur yang dianggap mewakili dalam pengelompokan dokumen proposal PkM.

Daftar Pustaka

[1] Ardanu, F., Himawan, H., & P, D. B. (2013). Pemanfaatan Teknologi Data Mining Dalam Menentukan Efektifitas Penyebaran Brosur.

[2] Dharmayanti, D., Bachtiar, A. M., & Heryandi, A. (2013). Pemodelan Data Warehouse, *12*(2), 151–168.

[3] Fadilah, U., Winarno, W. W., Amborowati, A., Fadilah, U., Winarno, W. W., & Amborowati, A. (2016). Perancangan Data Warehouse Untuk Sistem Akademik STMIK Kadiri Data Warehouse System Design For Academic STMIK Kadiri, *6*(2), 217–228.

[4] Ilmiah, J., Komputa, I., Volume, E., Issn, F., Cv, D. I., Anugerah, K., … Bandung, U. (2016a). PEMBANGUNAN PERANGKAT LUNAK DATA WAREHOUSE Jurnal Ilmiah Komputer Dan Informatika (KOMPUTA), *1*.

[5] Ilmiah, J., Komputa, I., Volume, E., Issn, F., Cv, D. I., Anugerah, K., … Bandung, U. (2016b). PEMBANGUNAN PERANGKAT LUNAK DATA WAREHOUSE Jurnal Ilmiah Komputer Dan Informatika (KOMPUTA). Ok

[6] Mulyati, S., Amini, S., & Juliasari, N. (2014). 104-279-1-PB.Pdf. *Jurnal Telematika MKOM, 6 No.1*.

[7] Ponniah, P. (2001). Data Warehouse Fundamentals: A Comprehensive Guide For IT Professional.J.Wiley. New York.

[8] M. Ainiyah, D. Nurmalasari, And W. Nengsih, "Visualisasi Data Teks Food Reviews Menggunakan Frequent Itemset Mining," J. Aksara Komput. Terap., Vol. 6, No. 2, 2017.

[9] L. Tanjaya, A. Wibowo, And D. Nurmalasari, "Sistem Pengelompokan E-Journal Berdasarkan Abstrak Menggunakan Text Mining Dan K-Means Clustering," J. Aksara Komput. Terap., Vol. 5, No. 1, 2016

[10] J. Han, J. Pei, And M. Kamber, Data Mining: Concepts And Techniques. Elsevier, 2011.

[11] R. Feldman And J. Sanger, The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data. Cambridge University Press, 2007.

[12] E. Muningsih, H. M. Nur, F. F. D. Imaniawan, V. R. Handayani, And F. Endiarto, "Comparative Analysis On Dimension Reduction Algorithm Of Principal Component Analysis And Singular Value Decomposition For Clustering," In Journal Of Physics: Conference Series, 2020, Vol. 1641, No. 1, P. 012101.

[13] A. Sukma, B. Zaman, And E. Purwanti, "Information Retrieval Document Classification With K-Nearest Neighbor," Rec. Libr. J., Vol. 1, No. 2, Pp. 129–138, 2015.

[14] A. N. Asyfa, D. Nurmalasari, And R. P. Sari, "Identifikasi Kinerja Perusahaan Berdasarkan Laporan Keuangan Menggunakan Algoritma K-NN," J. Aksara Komput. Terap., Vol. 5, No. 1, 2016.

[15] S. H. Myaeng, K. S. Han, And H. C. Rim, "Some Effective Techniques For Naive Bayes Text Classification," IEEE Trans. Knowl. Data Eng., Vol. 18, No. 11, Pp. 1457–1466, 2006

[16] W. Zhang And F. Gao, "Performance Analysis And Improvement Of Naïve Bayes In Text Classification Application," 2013 IEEE Conf. Anthol. Nthol. 2013, Pp. 1–4, 2013.

[17] M. A. Fauzi, S. Gosario, A. Z. Arifin, And I. S. Prabowo, "Klasifikasi Berita Berbahasa Indonesia Menggunakan Seleksi Fitur Dua Tahap Dan Naive Bayes," SYSTEMIC, Vol. 03, No. 02, Pp. 7–12, 2017.

[18] Nurhuda, F., Widya Sihwi, S. Dan Doewes, A. (2016) "Analisis Sentimen Masyarakat Terhadap Calon Presiden Indonesia 2014 Berdasarkan Opini Dari Twitter Menggunakan Metode Naive Bayes Classifier," Jurnal Teknologi & Informasi Itsmart, 2(2), Hal. 35