BAB III

METODOLOGI PENELITIAN

3.1. Gambaran Umum Objek Penelitian

Gambar 3.1. Logo Universitas Multimedia Nusantara

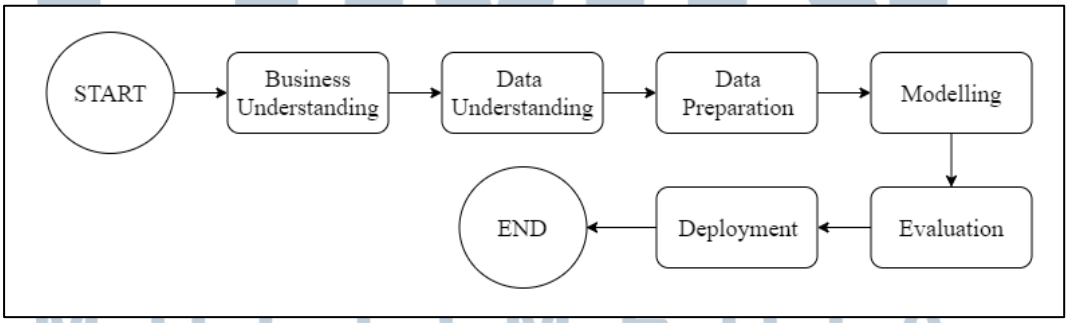
Sumber : Universitas Multimedia Nusantara

Objek dari penelitian ini terdapat dalam institusi pendidikan tinggi Universitas Multimedia Nusantara. Universitas Multimedia Nusantara atau disingkat dengan UMN secara resmi berdiri pada tanggal 25 November 2005 yang berada didalam naungan Yayasan Kompas Gramedia. Berlokasi di Kota Tangerang, Banten UMN menawarkan empat fakultas yaitu Fakultas Teknik & Informatika, Fakultas Seni & Desain, Fakultas Ilmu Komunikasi, Fakultas Bisnis, Fakultas Perhotelan dan juga menawarkan program studi bertaraf international. UMN juga memiliki beragam jenis program studi seperti Sistem Informasi, Teknik Informatika, Sistem Komputer, Teknik Fisika, Teknik Elektro, Manajemen, Akuntansi, Komunikasi Strategis, Jurnalistik, Desain Komunikasi Visual, Arsitektur, Film dan Animasi. Dengan beragam

cabang Pendidikan yang ditawarkan, UMN memiliki 198 dosen untuk memfasilitasi pengajaran sesuai bidang yang diampu.

UMN sebagai institusi Pendidikan tinggi tak lepas dari hubungan dengan insitusi lainnya, baik dari bidang Pendidikan maupun sektor lainnya seperti industri dan pemerintahan. UMN mendirikan Lembaga Penelitian dan Pengabdian Masyarakat atau disingkat sebagai LPPM, merupakan lembaga yang menangani bidang penelitian dan pengembangan ilmu, teknologi dan seni serta pengabdian masyarakat yang dijalankan oleh pihak UMN secara independen maupun berupa kerjasama dengan insitusi eksternal dari bidang pendidikan, industri, atau pihak pemerintahan. LPPM menangani tawaran kerjasama penelitian yang diberikan terhadap UMN dari pihak eksternal tersebut. Tawaran kerjasama penelitian nantinya akan dikumpulkan dan dialokasikan terhadap dosen – dosen di UMN berdasarkan bidang kompetensi dosen tersebut.

3.2. Metode Penelitian



Gambar 3.2. *Flowchart* Penelitian

Metode dalam penelitian ini berdasarkan pada penggunaan metode CRISP-DM seperti yang digambarkan dalam Gambar 3.2 dengan penjelasan sebagai berikut:

1. *Business Understanding*

   Sebelum memulai projek penelitian klasifikasi dan visualisasi, kebutuhan atau *requirements* harus dikumpulkan dan disusun secara matang dan lengkap. Hal ini dilakukan dengan melaksanakan proses wawancara dan diskusi dengan calon *user* yang akan memanfaatkan hasil dari penelitian. Dalam penelitian ini, calon *user* adalah pihak dari LPPM Universitas Multimedia Nusantara. Kebutuhan dari visualisasi, batasan dari informasi yang digunakan, aturan dalam menyusun visualisasi, dan hal – hal lainnya menjadi bahan wawancara dan diskusi dengan pihak LPPM Universitas Multimedia Nusantara. Dari wawancara tersebut akan dicari tujuan dari projek penelitian dan kebutuhan – kebutuhannya. Hasil dari wawancara ini akan disimpulkan ke dalam Tujuan dan Rumusan Masalah dalam penelitian.

2. *Data Understanding*

   Dalam penelitian ini jenis data yang digunakan terbagi atas 2 jenis yaitu Data Primer dan Data Sekunder.

   Jenis pertama yaitu merupakan data primer, atau data yang diperoleh langsung dari sumbernya baik melalui proses observasi

atau wawancara. Dalam penelitian ini menggunakan 2 data diantaranya data kompetensi dosen Universitas Multimedia Nusantara dan data tawaran kerjasama penelitian melalui proses wawancara dengan pihak LPPM Universitas Multimedia Nusantara.

Jenis kedua yaitu merupakan data sekunder, yaitu studi literatur dimana dilakukan riset dan studi terkait dengan teori dan metode yang digunakan selama penelitian, seperti *Text mining, Naïve Bayes Classifier*, *Data visualization*, dan sebagainya. Referensi yang digunakan oleh penelitian ini adalah buku – buku dan jurnal penelitian baik dalam bentuk fisik maupun digital yang memuat topik terkait dengan penelitian ini.

3. *Data Preparation*

Dalam tahap ini data – data dari tahap sebelumnya akan disiapkan untuk diolah ke dalam proses klasifikasi. Data kompetensi dosen dan data tawaran kerjasama penelitian akan disimpan dalam satu *database*. Hal ini dilakukan supaya data tersimpan secara terpusat dan aman. Data yang disimpan dalam *database* juga mudah untuk digunakan nantinya dalam tahap klasifikasi dan visualisasi. Untuk berpindah dari format data *spreadsheet* dan *pdf*, data kompetensi dosen dan data tawaran kerjasama penelitian akan dimasukkan ke dalam *database* menggunakan aplikasi berbasis *web* yang mempunyai fitur untuk memanipulasi data (*create, read,*

delete, dan *update*). Dalam proses memasukkan data ke dalam *database* baik data kompetensi dosen maupun data tawaran kerjasama penelitian, juga dilakukan proses *data cleansing*, dimana kesalahan penulisan dalam data dan tanda baca yang tidak diperlukan akan dihapus dan diperbaiki. Setelah kedua data selesai dimasukkan ke dalam *database,* dari atribut – atribut data kompetensi dosen akan ditambahkan satu atribut yang ditentukan sebagai *label* untuk proses klasifikasi. Atribut ini bernama *Major*, yaitu berisikan kategori kompetensi dosen yang merujuk kepada kategori – kategori yang terdapat dalam Rencana Induk Penelitian LPPM Universitas Multimedia Nusantara.

Dalam klasifikasi dibutuhkan *training dataset* yang berguna sebagai basis dari klasifikasi yang akan digunakan oleh algoritma saat model sedang diaplikasikan ke dalam *dataset* yang akan diklasifikasi. Sebelum diaplikasikan model, *training dataset* dipersiapkan dengan proses dalam prosedur *text processing* yaitu *Tokenizing* yaitu memisahkan kalimat menjadi teks – teks yang independent, *stemming* yaitu proses perubahan kata – kata dalam dokumen menjadi bentuk dasar dari kata tersebut dengan cara menghilangkan imbuhan, awalan, dan akhiran dari kata, mengganti ukuran huruf, dan menghitung *n-gram* dimana *n-gram* merupakan kondisi diberikan dalam sebuah paragraf dengan mengisi nilai dari *n* dimana nilai dari *n* sama dengan jumlah kumpulan sukukata, huruf,

maupun kata, seperti contoh dengan kalimat “Ibu Budi merupakan penjaja makanan ringan” denga nilai n-gram = 2 (atau disebut juga dengan bigram) maka akan menghasilkan kumpulan kata sebagai berikut : “Ibu Budi”, “Budi merupakan”, “merupakan penjaja”, dan seterusnya. Dapat disimpulkan bahwa nilai *x* adalah jumlah kata dalam suatu kalimat A, maka jumlah *n-gram* dari kalimat A bias dicari menggunakan rumus *n-gram* A = $x - (n - 1)$.

Setelah seluruh proses ini selesai dilakukan maka akan menghasilkan *dataset* berupa *word list* yang siap untuk diklasifikasi menggunakan model klasifikasi.

4. *Modelling*

Untuk membentuk model klasifikasi sebelumnya dibutuhkan *training dataset. Training dataset* yang dimaksud merupakan data kompetensi dosen namun data tersebut sudah dilengkapi dengan atribut *Major,* yaitu *label* yang sudah benar untuk masing – masing dosen. *Training dataset* lalu diaplikasikan dengan operator klasifikasi *Naïve Bayes Classifier* sehingga membentuk model klasifikasi yang utuh. Langkah selanjutnya yang dilakukan adalah menyambungkan data kompetensi dosen yang belum diklasfikasi terhadap model klasifikasi tersebut untuk menyelesaikan tahap klasifikasi.

5. *Evaluation*

Hasil dari klasifikasi data kompetensi dosen lalu dievaluasi dan diteliti. Tingkat kesuksesan dari model klasifikasi *Naïve Bayes Classifier* dinilai dari tingkat akurasi dalam *confusion matrix* yang dihasilkan. Dalam tahap ini juga di evaluasi apakah hasil klasifikasi tersebut sudah memenuhi kebutuhan dari tahapan *Business Understanding*. Proses validasi secara manual pun dilakukan dengan membandingkan hasil klasifikasi dengan data kompetensi dosen yang memiliki *label* yang sudah dipastikan benar untuk melihat tingkat akurasi model klasifikasi terhadap data yang benar.

6. *Deployment*

Hasil dari klasifikasi dimasukkan ke dalam *database MySQL* untuk digunakan oleh visualisasi yang nantinya akan dibandingkan dengan data tawaran kerjasama penelitian.

Microsoft Power BI digunakan untuk mengolah data ke dalam bentuk visualisasi data. Dalam tahap ini kedua data yaitu data tawaran kerjasama penelitian dan data kompetensi dosen ditampilkan. *Parameter - parameter* yang digunakan dalam visualisasi disesuaikan dengan hasil wawancara dari tahap *Business Understanding*.

3.3. Metode Klasifikasi

Ada 2 metode klasifikasi yang dipertimbangkan untuk digunakan dalam penelitian ini, yaitu sebagai berikut :

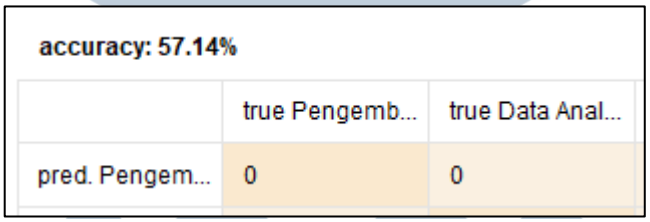
Tabel 3.1. Sifat Metode – Metode Klasifikasi

Naïve Bayes Classifier	C.45 Algorithm
Training dataset harus memiliki *label* yang akan dijadikan hasil klasifikasi terhadap data baru	Mengklasifikasikan data berdasarkan aturan – aturan yang diterapkan terhadap atribut data untuk memutuskan hasil klasifikasi
Mempelajari frekuensi dari atribut – atribut dalam *training dataset*	Membentuk *"*cabang*"* dari pohon keputusan (*decision tree*) berdasarkan atribut yang dipilih sebagai "akar"
Mengklasifikasikan data berdasarkan *label* yang didapat dari atribut – atribut yang dimiliki data	Menggunakan *Entropy* dan *Information Gain* sebagai acuan

Secara singkat, *Naïve Bayes Classifier* adalah metode klasifikasi probabilistik yang menghitung probabilitas dengan cara mencari dan menghitung frekuensi dan kombinasi atribut dalam data yang diberikan. Metode tersebut menggunakan *training dataset*, yaitu data dengan atribut – atribut dan *label* yang benar yang digunakan untuk mempelajari atribut – atribut yang dimiliki data tersebut, lalu dibandingkan dengan atribut – atribut dalam data yang akan diklasifikasi.

Metode *C4.5 Algorithm* merupakan salah satu varian dari metode *decision* tree, dengan cara kerja yaitu membentuk pohon keputusan dengan memilih atribut yang memiliki nilai (*weight*) yang paling besar atau dominan sebagai "akar", lalu membentuk "cabang" dari nilai atribut tersebut dan mengisi "daun" tersebut dengan hasil dari klasifikasi. Proses pembentukan

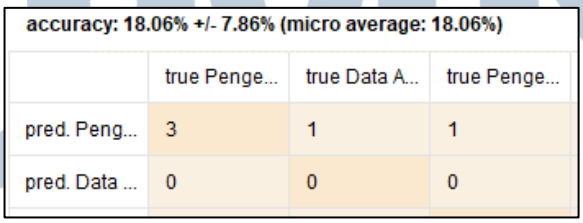
"cabang" tersebut diulangi hingga setiap kemungkinan dari nilai atribut terisi. Metode ini mengandalkan *Entropy* (probabilistik dari suatu kejadian) dan *Information Gain* (mengukur seberapa banyak informasi yang diberikan dari suatu atribut) untuk menguji metode tersebut. Perbandingan kedua metode ini dapat dilihat dalam Tabel 3.1. Untuk menguji dan memilih metode klasifikasi yang tepat, peneltian ini menggunakan *training dataset* yang menggunakan basis data kompetensi dosen sebagai data yang akan diuji coba terhadap tiap – tiap metode klasifikasi. Dalam Gambar 3.3. dan Gambar 3.4. merupakan hasil dari tingkat akurasi masing – masing metode menggunakan metode *confusion matrix*.



accuracy: 57.14%

	true Pengemb...	true Data Anal...
pred. Pengem...	0	0

Gambar 3.3. Tingkat Akurasi untuk *Naïve Bayes*



accuracy: 18.06% +/- 7.86% (micro average: 18.06%)

	true Penge...	true Data A...	true Penge...
pred. Peng...	3	1	1
pred. Data ...	0	0	0

Gambar 3.4. Tingkat Akurasi untuk C4.5

Untuk mencapai hasil penelitian yang optimal, metode *Naïve Bayes* akan digunakan sebagai metode klasifikasi dalam penelitian ini dikarenakan penggunaan metode tersebut memiliki tingkat akurasi yang lebih tinggi.

3.4. Spesifikasi Sistem

Penelitian ini menggunakan beberapa perangkat keras dan lunak yang dijabarkan sebagai berikut:

3.4.1. Perangkat Keras

CPU : Intel i5-4670 3.4 GHz

RAM : 8 GB DDR3

GPU : NVIDIA GTX 1060 6GB

SSD : 250GB 850 EVO

3.4.2. Perangkat Lunak

- *Tools* untuk *data preprocessing*, *modelling*, dan *classification*: RapidMiner dan Microsoft Excel. Komparasi *tools* untuk proses klasifikasi dilakukan antara program RapidMiner dan WEKA. Alasan utama untuk penggunaan program RapidMiner adalah fitur Operator (fungsi – fungsi yang terdapat dalam program di dalam bentuk blok yang disambung menggunakan *port input* dan *output*) yang mencakup lebih banyak proses *data mining* dibandingkan WEKA. Hal ini mendukung fleksibilitas dan kemudahan dalam tahap *Data Preparation* dan *Modelling*. RapidMiner juga dapat memiliki memori yang lebih besar sehingga dapat memroses *dataset* yang lebih besar dibandingkan dengan WEKA (Foozy, Ahad, Abdollah, & Wen, 2017). Microsoft Excel digunakan untuk melihat dan meninjau data kompetensi dosen sebelum dimuat ke dalam *database*.

- *Tools* untuk pengembangan aplikasi pendukung: XAMPP, Brackets, Google Chrome, Mozilla Firefox. XAMPP dan Brackets digunakan untuk mengembangkan aplikasi *database* berbasis *web*. Dikarenakan aplikasi hanya dijalankan secara lokal (*localhost*) maka digunakan XAMPP yang dapat menjalankan Apache Web Server secara lokal, Brackets digunakan sebagai *text editor* dalam pemrograman PHP. Fitur *Live Edit* yang dimiliki Brackets juga memudahkan dalam tahap *troubleshooting* aplikasi. *Internet browser* seperti Mozilla Firefox dan Google Chrome digunakan sebagai media *testing* aplikasi dan untuk penggunaan aplikasi.
- *Tools* untuk *data visualization*: Microsoft Power BI. Komparasi *tools* untuk visualisasi dipertimbangkan antara Microsoft Power BI dan aplikasi visualisasi Tableau. Alasan utama digunakannya Power BI dikarenakan Power BI dapat mendukung *data source* yang sangat beragam, terutama dari Microsoft Office 365 (Gowthami & Kumar, 2017). Microsoft Power BI juga lebih cocok digunakan dalam *deployment* visualisasi yang bersifat *straightforward* dan sederhana, hal ini dikarenakan Microsoft Power BI juga lebih ditujukan ke pasar dengan *beginner* dan *novice users*, berbeda halnya dengan Tableau yang ditujukan untuk pasar *power users* yang memiliki infrastruktur data yang lebih besar (*datawarehouse*). Lebih banyak beragam dukungan dan parameter untuk visualisasi yang beragam juga menjadi

alasan digunakannya Power BI. Dalam implementasi, biaya produk Microsoft Power BI juga jauh lebih murah dibandingkan dengan biaya Tableau.

3.5. Penelitian Terdahulu

Dalam projek penelitian ini telah digunakan 3 penelitian terdahulu yang digunakan sebagai acuan dan referensi dalam mengerjakan dan menyelesaikan projek penelitian. Penelitian terdahulu pertama yang berjudul "*Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks"* digunakan dalam projek penelitian ini sebagai referensi dalam proses penerapan metode *Naïve Bayes Classifier* dalam mengklasifikasi data. Penelitian terdahulu kedua yang berjudul "*Text and Image Based Spam Email Classification using KNN, Naive Bayes and Reverse DBSCAN Algorithm"* digunakan dalam projek penelitian ini sebagai acuan dalam bagaimana metode *Naïve Bayes* unggul sebesar 87% dibandingkan dengan metode lainnya dalam sutdi kasus klasifikasi teks. Dan penelitian terdahulu ketiga yang berjudul "*Modified Approach of Multinomial Naïve Bayes for Text Document Classification*" digunakan dalam projek penelitian ini sebagai acuan dalam bagaimana proses klasifikasi teks berita menggunakan metode *Naive Bayes*, yang secara garis besar merupakan salah satu tahapan dari projek penelitian. Tabel 3.2. menjabarkan penelitian terdahulu yang digunakan dalam projek penelitian ini.

Tabel 3.2. Penelitian Terdahulu

No	Nama Peneliti	Tahun Dirilis	Penelitian	Hasil Penelitian
1.	Dewan Md. Farid, Li Zhang, Chowdhury Mofizur Rahman, M.A. Hossain, Rebecca Strachan	2014	*Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks*. Expert Systems with Applications 41, 1937–1946	Metode klasifikasi yang diusulkan dalam penelitian (*Decision Tree* & *Naïve Bayes*) diuji menggunakan *classification accuracy*, *precision*, *sensitivity specificity analysis*, dan *10-fold cross validation* menggunakan 10 *dataset* uji berbeda yang tersedia di UCI Machine Learning Depository. Hasil penelitian menunjukkan metode yang diusulkan sangat memuaskan untuk klasifikasi *multiclass* dalam kehidupan nyata. Penelitian ini menjadi panduan dalam proses penerapan metode *Naïve Bayes Classifier* untuk tujuan klasifikasi.
2.	Anirudh Harisinghaney, Arnan Dixit, Saurabh Gupta, Anuja Arora	2014	*Text and Image Based Spam Email Classification using KNN, Naive Bayes and Reverse DBSCAN Algorithm*. 2014 International Conference on	Penelitian ini membandingkan 3 metode berbeda (k-NN, Naïve Bayes, Reverse DBSCAN) untuk keperluan deteksi *spam email*. Perbandingan kedua metode klasifikasi yaitu k-*Nearest*

			Reliability, Optimization and Information Technology - ICROIT 2014, 153-155	*Neighbor* atau k-NN dengan *Naïve Bayes* menjadi panduan dalam pemilihan metode klasifikasi data kompetensi dosen. Hasil penelitian ini menunjukkan bahwa metode *Naïve Bayes* memiliki tingkat akurasi sebesar 87%, lebih tinggi dibandingkan dengan tingkat akurasi oleh metode k-NN yaitu sebesar 83%.
3	S.W. Mohod, Dr. C.A. Dhote, Dr. V.M. Thakare	2015	*Modified Approach of Multinomial Naïve Bayes for Text Document Classification*. 2015 International Journal of Computer Science & Communication – IJCSC 2015, 196-200	Penelitian ini menggunakan metode *Multinomial Naïve Bayes* dengan sedikit modifikasi untuk keperluan proses klasifikasi teks menggunakan *dataset* berita Reuters 21578. Hasil penelitian menunjukkan peningkatan tingkat akurasi hasil klasifikasi jika dibandingkan dengan pemakaian metode *Multinomial Naïve Bayes* tanpa modifikasi.