BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

3.1.1 CLICK-ID

Dataset CLICK-ID adalah kumpulan headline berita Indonesia yang dikumpulkan dari 12 penerbit berita online lokal. detikNews, Fimela, Kapanlagi, Kompas, Liputan6, Okezone, PosmetroMedan, Republika, Sindonews, Tempo, Tribunnews, Wowkeren. Dataset ini terdiri dari dua bagian. 46.119 data artikel mentah dan 15.000 contoh judul beranotasi clickbait. Anotasi dibuat oleh tiga annotator yang memeriksa setiap judul. Putusan hanya didasarkan pada judul. Dalam hal ini, mayoritas dianggap sebagai kebenaran dasar. Dalam sampel beranotasi, anotasi menunjukkan 6.290 clickbait dan 8.710 non-clickbait.

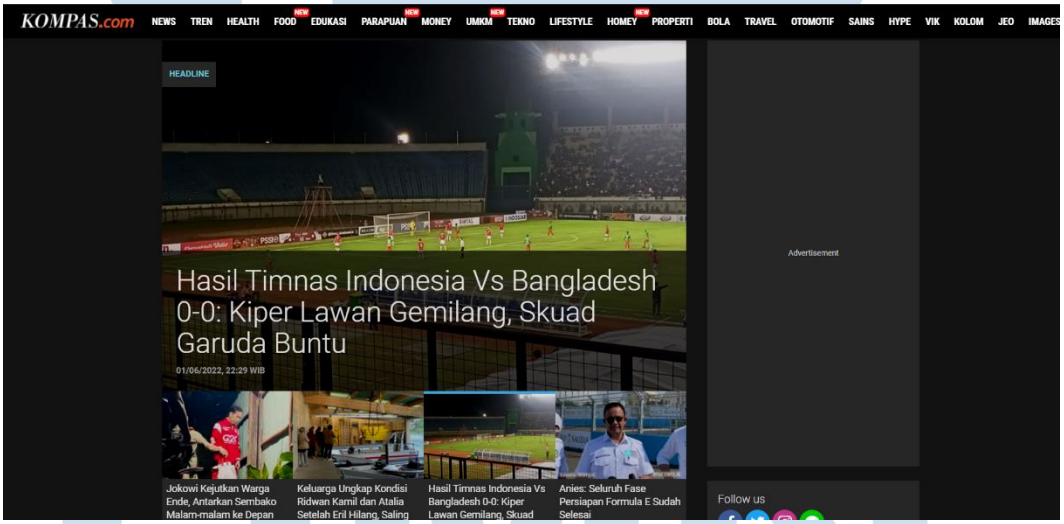
3.1.2 Kompas

Kompas.com adalah salah satu situs berita terkemuka di Indonesia yang menyediakan berbagai informasi terkini tentang berita nasional, internasional, ekonomi, politik, hiburan, olahraga, dan banyak lagi. Sebagai platform berita online, Kompas.com berfokus pada menyajikan berita terbaru dengan beragam perspektif yang objektif, terpercaya, dan berimbang.

Kompas.com didirikan pada tahun 1995 dan merupakan bagian dari Kompas Gramedia, salah satu kelompok media terbesar di Indonesia. Situs ini memanfaatkan teknologi internet untuk menyebarkan berita secara luas kepada masyarakat Indonesia dan dunia. Kompas.com hadir dengan antarmuka yang user-friendly, memudahkan pembaca untuk menavigasi dan mengakses berbagai konten berita yang disajikan.

Ketika mengunjungi Kompas.com, pembaca akan melihat tampilan beranda yang menampilkan judul-judul berita terkini dari berbagai kategori. Terdapat juga gambar utama yang menarik perhatian pembaca dan memberikan gambaran visual terkait berita tersebut. Di bagian atas beranda, terdapat menu navigasi yang memudahkan pembaca untuk menjelajahi kategori berita yang diminati.

Setiap berita yang disajikan di Kompas.com dilengkapi dengan teks, gambar, dan video pendukung yang memberikan informasi secara komprehensif. Kompas.com juga sering kali melampirkan link-link terkait atau berita terkait lainnya, sehingga pembaca dapat melihat berbagai sudut pandang atau informasi tambahan terkait topik yang sedang dibahas.



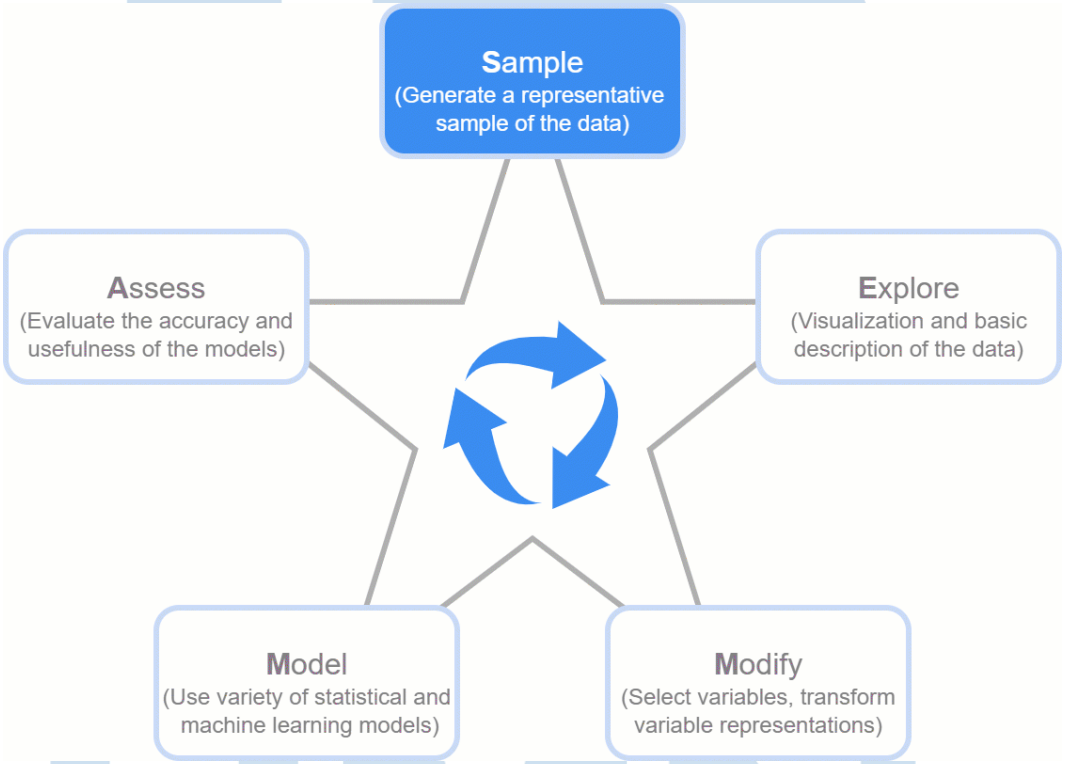
Gambar 3.1 Kompas.com

Kompas mencapai puncak oplah pada tahun 2004, ketika sirkulasi harian mencapai sekitar 530.000 eksemplar, dan edisi Minggu, 610.000 eksemplar. Jumlah pembaca mencapai sekitar 2,25 juta. Pada tahun 2014 oplahnya mencapai 507.000, dengan 66% beredar di Jabodetabek[15].

3.2 Metode Penelitian

Penelitian dilakukan dengan menggunakan metodologi *SEMMA*, yaitu metodologi *data science* yang dikembangkan oleh SAS Industry untuk mengubah suatu data menjadi Informasi.[16]

Proses *SEMMA* merupakan seperti berikut:



Gambar 3.2 SEMMA Methodology[17]

Metode ini terdiri dari 5 proses yaitu :

- *Sample*: Mengambil Sampel dari dataset
- *Explore* : Menelusuri dataset yang sudah dilakukan *sampling*
- *Modify* : Memilih dan mengubah variabel
- *Model* : Membuat modelling, dengan kasus ini menggunakan *Machine Learning*
- *Assess:* Mengevaluasi akurasi dan membandingkan model mana yang paling berguna.

Tabel 3.1 SEMMA vs CRISP-DM

SEMMA	CRISP-DM
--	Business Understanding
Sample	Data Understanding
Explore	
Modify	Data Preparation
Model	Modeling
Assessment	Evaluation
--	Deployment
Used for Data Science	Used for E-Commerce

Berdasarkan tabel perbandingan, metodologi *SEMMA* lebih dipilih dikarenakan SEMMA digunakan untuk melakukan *Data Science*, walaupun metode seperti *CRISP-DM* dan *SEMMA* memiliki fungsi yang sama, *CRISP-DM* lebih mengutamakan pengaruh Informasi terhadap bisnis E-Commerce, dan SEMMA menggunakan informasi tersebut untuk membuat model *data science*, *SEMMA* memiliki proses yang lebih *simple* dan juga terfokus dalam mengolah *data* menjadi *informasi* sehingga cocok dalam pembuatan projek ini.

3.3 Teknik Pengumpulan Data

Data dikumpulkan melalui Teknik Data Scrapping, yaitu metode di mana peneliti menggunakan program komputer untuk mengekstrak data dari output program lain. Dalam konteks penelitian ini, peneliti memanfaatkan teknik ini untuk mengumpulkan data dari situs web Kompas yang memiliki headline populer di Indonesia. Kompas memiliki volume yang sangat besar dan menyediakan berbagai macam artikel, laporan, dan juga konten berita setiap harinya, serta mencakup topik yang cukup diversifikasi untuk mendapatkan berbagai topik untuk dijadikan kata kunci yang akan di latih oleh algoritma[18][19].

Proses pengumpulan data dilakukan dengan memasuki situs web Kompas dan mengambil judul-judul headline yang akan dianalisis. Peneliti menggunakan bahasa pemrograman *Python* dengan *library* yang bernama

beautifulsoup4 yang memungkinkan untuk mengakses dan mengekstrak informasi ini secara otomatis. Data yang dikumpulkan berasal dari indeks situs web Kompas, dan dalam penelitian ini, sebanyak lebih dari 9000 judul headline berhasil diambil serta waktu dan tanggal berita tersebut di *publish*, waktu data di kumpulkan berdurasi lebih dari 6 jam total, dengan delay data scrapping 5-10 detik untuk menghindari potensi koneksi diputus karena dianggap sebagai serangan *denial-of-service* (DoS).

Setelah proses pengumpulan data selesai, data tersebut kemudian digabungkan menjadi satu dataset. Hal ini dilakukan untuk memperbarui dataset publik CLICK-ID yang sebelumnya terdiri dari 15.000 sampel data. Dengan penambahan data baru dari pengumpulan ini, total jumlah data dalam dataset CLICK-ID menjadi ±24.000.

Penggunaan teknik Data Scrapping dalam penelitian ini memungkinkan peneliti untuk mengakses informasi yang relevan dengan topik penelitian secara efisien. Dalam hal ini, fokus penelitian adalah pada analisis judul-judul headline, dan data yang dikumpulkan melalui Data Scrapping memberikan akses langsung ke informasi tersebut.

Pengumpulan data yang dilakukan melalui teknik Data Scrapping juga memiliki keuntungan dalam hal skala dan waktu. Dengan menggunakan program komputer untuk melakukan pengumpulan data secara otomatis, peneliti dapat mengumpulkan data dalam jumlah yang signifikan dalam waktu yang relatif singkat

Tabel 3.2 CLICK-ID dataset distribution headlines and clickbait titles

Publishers	Artikels	Non-Annotated	annotated		
			total	non-clickbait	clickbait
detikNews	5468	4468	1000	890	110
fimela	788	88	700	306	394
kapanlagi	1006	6	1000	603	397
kompas	3243	1743	1500	1157	343
liputan6	4581	3081	1500	613	887
okezone	4664	3164	1500	741	759
posmetro	307	7	300	71	229
republika	5782	4282	1500	1267	223
sindonews	3572	2072	1500	1215	285
tempo	4026	2526	1500	1118	382
tribunnews	9662	8162	1500	451	1049
wowkeren	3020	1520	1500	278	1222
Total	46119	31119	15000	8710	6280

Tabel 3.3 Dataset Scrapping Kompas Top 9

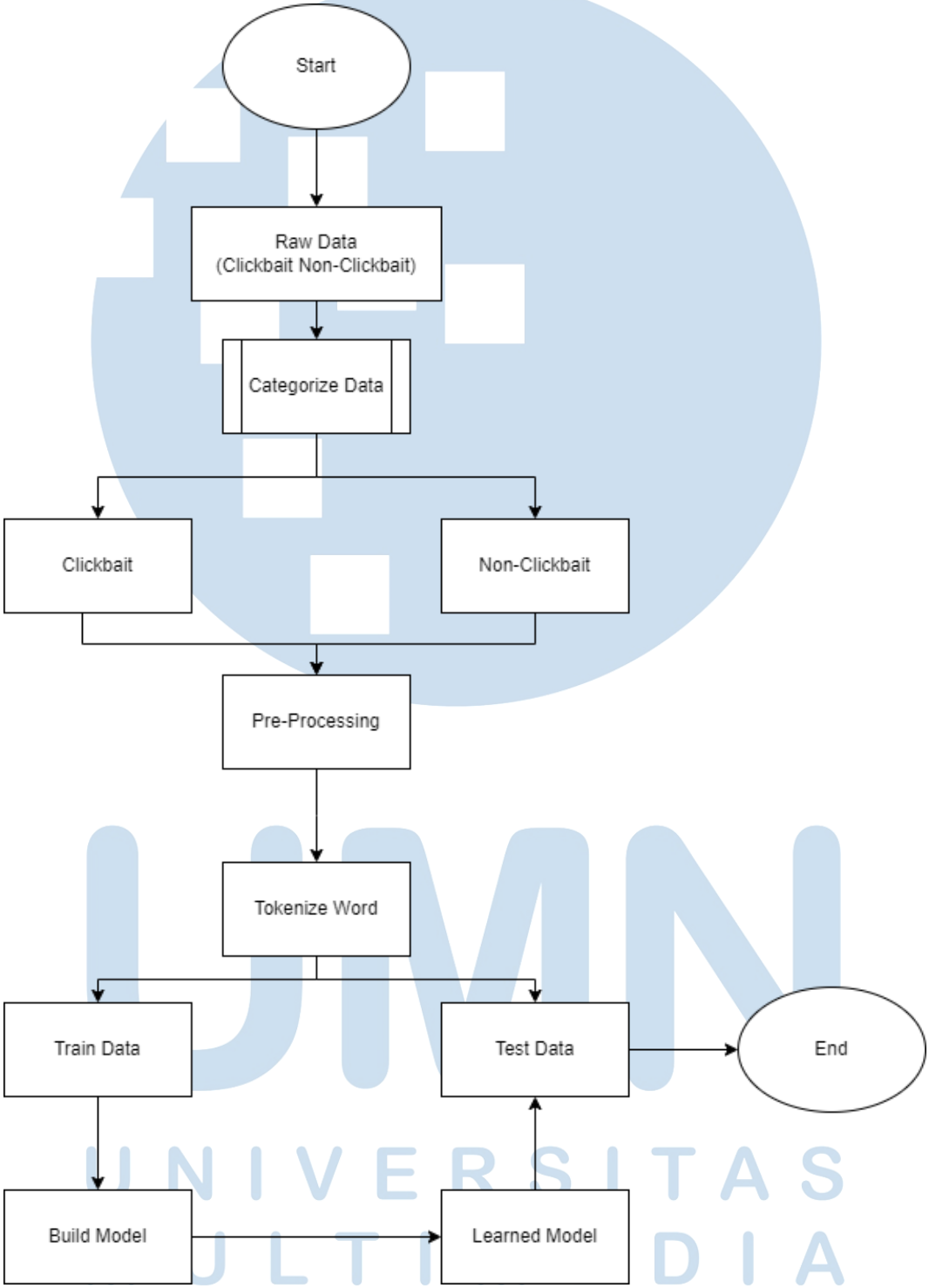
title	date
Posisi Strategis Sandiaga di PPP Diumumkan Setelah Rapimnas	6/12/2023 10:46
Rencana Pertemuan Puan-AHY serta Jejak Rivalitas PDI-P dan Demokrat	6/12/2023 10:43
Kemenag: Layanan Katering untuk Jemaah Haji di Mekkah Berhenti Sementara pada 7, 14, dan 15 Zulhijjah	6/12/2023 10:38
Sinyal PPP Usung Sandiaga Uno Jadi Cawapres Pilpres 2024...	6/12/2023 10:23
MA Tolak Gugatan PSI Terkait Aturan Pendirian Rumah Ibadah	6/12/2023 9:56
KPK Sebut Lukas Enembe Akan Jalani Sidang Secara "Online" Hari Ini	6/12/2023 9:32
Blak-blakan Mahfud MD Bicara Transaksi Balik Meja DPR dan Penyusup di Penegak Hukum	6/12/2023 9:21
Eksaminasi Akademisi Atas Putusan Sambo: Pasal Pembunuhan Berencana Dinilai Kurang Tepat Digunakan	6/12/2023 9:19
Hari Ini, Sidang Perdana Praperadilan Sekretaris Nonaktif MA Hasbi Hasan Digelar	6/12/2023 8:41

3.4 Variabel Penelitian

Variabel yang ada di dalam Penelitian merupakan:

- Variabel Independent : Merupakan varabel data penyebab dan tidak ada hubungannya dengan variable lain, variable tersebut merupakan *data Headlines* yang sudah dilakukan *Tokenization* dalam *pre-processing*.
- Variabel Dependent : Merupakan hasil yang didapatkan dari Variabel Independent, yaitu akurasi kemampuan algoritma *Machine Learning* dalam mendeteksi apakah *Headlines* yang akan diinput dapat dikenal sebagai *Clickbait* atau bukan.

3.5 Alur Penelitian



Gambar 3.5.1 Proposal Workflow

Berikut merupakan desain *workflow* yang akan dilakukan oleh *Machine Learning* yang dipilih/dibandingkan sebelum di implementasikan, dimulai dengan data mentah yang akan diambil menggunakan teknik *data scrapping*, kemudian data tersebut akan dikategorikan sebagai judul *clickbait* dan *non-clickbait*, untuk di *pre-processing* menjadi *token* dan dimasukkan kedalam *Machine Learning* untuk dilatih dan diuji akurasinya, bagian ini akan terus diulang sampai mendapatkan akurasi yang dapat mendeteksi secara konsisten.

3.6 Teknik Analisis Data

3.6.1 Kuantitatif

Penelitian kuantitatif didefinisikan sebagai studi sistematis tentang fenomena dengan mengumpulkan data kuantitatif dan menerapkan teknik statistik, matematika, atau komputasi. Penelitian kuantitatif mengumpulkan informasi dari pelanggan yang ada dan potensial melalui prosedur pengambilan sampel, survei online, survei online, survei, dan pengiriman lainnya. Hasilnya dapat digambarkan dalam format numerik. Setelah memahami angka-angka ini dengan cermat, antisipasi masa depan produk atau layanan Anda dan lakukan perubahan yang sesuai.

Dalam proses menentukan algoritma, bahasa pemrograman yang akan digunakan adalah bahasa pemrograman *Python*, dengan tingkat *library* yang luas seperti *matplotlib* dan juga *scikit* & *tensorflow*, dapat mengimplementasikan *machine learning* dengan mudah dibandingkan bahasa pemrograman lainnya, metode *machine learning* yang akan diimplementasikan adalah *Naïve Bayes* dan juga *Recurrent Neural Network*, kedua metode ini merupakan metode klasifikasi *statistical classification*.

Tabel 3.4 Naive Bayes vs Recurrent Neural Network

No.	Machine Learning		
	Comparison	Naïve Bayes	Recurrent Neural Network
1	Category	Generative	Discriminative
2	Library	Scikit-Learn, NaïveBayes	Tensorflow, Theano, Keras, Caffe

Seperti yang sudah ada di dalam tabel perbandingan, Naïve Bayes merupakan *generative model* dimana pada saat dilakukan *training*, NB akan mencoba untuk mengetahui bagaimana cara data tersebut bisa dihasilkan, dengan kata lain mencari distribusi data yang menghasilkan hasil yang didapat dari contoh input yang dimasukkan kedalam model. Untuk menggunakan Naïve Bayes, data pertama harus diubah dalam proses *pre-processing* ke dalam bentuk *vector* yang berisi nilai numerik yang akan dijadikan sebagai input pada model NB. *Classifier* ini akan mengasumsikan bahwa *features* (*attributes* dari *vector)* merupakan nilai *independent* dari satu sama lain.

Disisi lain *Recurrent Neural Network* merupakan *discriminative model*, dengan menggunakan *conditional probability* RNN akan menentukan perbedaan antara negatif dan positif dari input yang diberikan untuk melakukan klasifikasi [20]. RNN mengubah data dalam *pre-processing* menjadi *data sequential*, dengan ini RNN dapat menjaga *memory* atas apa yang sudah dibaca sebelumnya, menjadikan RNN sangat cocok untuk mengolah kalimat dikarenakan kemampuan RNN untuk mengkorelasikan data yang baru dengan data yang lama.