BAB II

LANDASAN TEORI

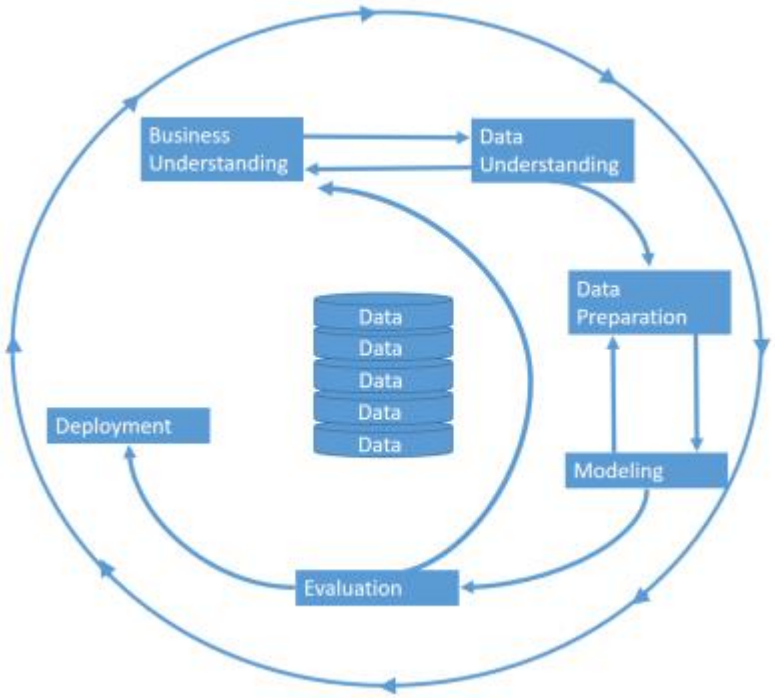
2.1 Pelecehan Seksual

Pelecehan seksual merupakan tindakan seksual yang tidak diinginkan oleh korban yang dilakukan oleh oknum. Pelecehan juga merupakan salah satu bentuk dari kekerasan seksual. Terdapat tiga dimensi dalam pelecehan seksual yaitu pelecehan *gender*, perhatian seksual yang tidak diinginkan, dan pemaksaan seksual. Pelecehan seksual sering terjadi di area perkotaan, kampus, tempat kerja, kantor atau tempat yang sepi dan korbannya merupakan seorang perempuan dan tidak jarang laki-laki yang menjadi korban akan tetapi pelaku yang melakukan tindakan pelecehan seksual adalah jenis kelamin laki-laki [12].

Dalam pelecehan seksual terbagi menjadi dua yaitu pelecehan seksual nonverbal dan verbal dimana nonverbal bearti pelecehan yang berkaitan dengan fisik seperti sentuhaan dan meraba biasanya dilakukan saat bertemu langsung dengan pelaku. Untuk pelecehan verbal bearti pelecehan yang tidak berkaitan dengan fisik yaitu menggunakan kata-kata biasanya dilakukan pada saat *chatting* [12].

2.2 CRISP-DM

CRISP-DM atau *Cross-Industry Standard Process for Data Mining* merupakan salah satu metode *framework* yang sudah standar *de factor* dalam dalam implementasi projek *data mining*. CRISP-DM memiliki siklus yang terdiri dari 6 tahapan atau fase yaitu *business understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment* [13].



Gambar 2. 1 Siklus CRISP-DM [9]

Pada gambar 2.1 merupakan proses siklus dari 6 fase atau tahapan yang terhubung dan dijelaskan pada berikut ini:

1. **Business Understanding**

   Business Understanding merupakan tahapan pertama dari siklus CRISP-DM yang bertujuan untuk mendapatkan gambaran tentang permasalahan pada bisnis dan inisialisasi kriteria keberhasilan dari projek yang ingin dicapai [13].

2. **Data Understanding**

   Tahapan kedua merupakan tahapan *Data Understanding* yang digunakan untuk mengumpulkan data, mengeksplorasi, mendeskripsikan, dan memeriksa kualitas data yang bertujuan untuk mengurangi potensi masalah dalam data serta mendapatkan wawasan atau pemahaman dari data yang sesuai dengan *Business Understanding* [13].

3. **Data Preparation**

Tahapan ketiga merupakan tahapan *Data Preparation* yang digunakan untuk memperbaiki dan membersihkan data yang telah ditemukan di tahapan *Data Understanding* serta melakukan beberapa proses yang dibutuhkan berdasarkan permasalahan utama dari tahapan 1 dan 2 dan juga tahapan ini akan ditinjau kembali dari tahapan *modelling* [13].

4. **Modeling**

   Tahapan keempat merupakan tahapan yang digunakan untuk memilih teknik atau algoritma pemodelan yang bergantung pada permasalahan bisnis serta jika terdapat permasalahan diharuskan untuk kembali ke tahapan *Data Preparation* [13].

5. **Evaluation**

   Tahapan kelima setelah melakukan melakukan *modelling*, akan di proses evaluasi kelayakan pada model yang telah dibuat terhadap tujuan utama pada permasalahan bisnis yang telah ditetapkan pada tahapan 1 [13].

6. **Deployment**

   Tahapan terakhir merupakan tahapan implementasi atau penerapan pada model yang telah dibuat ke dalam bisnis atau tujuan permasalahan yang telah di tetapkan [13].

2.3 Data Mining

Data Mining merupakan bidang ilmu komputer dengan menggunakan algoritma matematika yang dapat mengumpulkan informasi dengan skala yang besar yang digunakan untuk menemukan pola baru atau informasi yang belum diketahui sebelumnya secara otomatis dan efisien serta dapat digunakan untuk pengambilan keputusan. Data Mining telah digunakan di berbagai bidang seperti di sektor keuangan, telekomunikasi, asuransi dan ritel termasuk persetujuan pinjaman / kartu kredit , deteksi penipuan, segmentasi pasar, analisis tren, pemasaran yang lebih baik, analisis pembelanjaan, dll [14].

2.4 Text Mining

Text mining merupakan tipe khusus dari *data mining* yang dapat menambang data dari kumpulan teks yang banyak dan besar serta dapat mengekstraksi informasi dari data teks untuk menemukan informasi yang baru dan belum diketahui seperti informasi yang berupa pola dan relasi antar teks. *Teks mining* juga digunakan dalam melakukan *text classification*, *text clustering*, *information extraction*, *document summarization*, dan *opinion mining* atau sentimen analisis. *Text mining* melibatkan *natural language processing* yang dapat membantu analisa dan mengolah data teks atau juga disebut *text preprocessing*. Selain itu, *text preprocessing* juga dapat membantu masalah terkait bahasa yang tidak konsisten, penggunaan bahasa yang tidak tepat, penggunaan bahasa *slang*, perbedaan sintak atau bahasa khusus. *Teks mining* dapat membantu mengidentifikasi informasi tentang suatu sentimen yang bertujuan untuk mendeteksi sebuah masalah dan mendapatkan solusi [15].

2.5 Sentimen Analisis

Sentimen analisis merupakan mengidentifikasi dan mengekstrak sebuah data kedalam informasi dan dapat mengetahui yang mereka ungkapkan dan kemudian dilakukannya klasifikasi polaritas ke dalam kategori. Kategori sendiri terdiri dari negatif, positif dan netral. Sentimen analisis juga dapat membantu untuk memahami sentimen pada setiap situasi seperti sentimen publik terhadap suatu kondisi seperti ulas barang, pasar keuangan, hubungan antar pelanggan, strategi pemasaran, dll. Saat ini sentimen analisis memiliki kemampuan untuk menganalisis teks lebih lanjut dengan menggunakan *machine learning* dan *deep learning*. Dengan aspek tersebut dapat mengklasifikasikan lebih akurat dan mendalam. Terdapat tiga level sentimen yang dapat dilakukan yaitu tingkat dokumen, tingkat kalimat, dan tingkat aspek [16].

2.6 Natural Language Processing

Natural Language processing adalah bagian cabang dari ilmu komputer dan kecerdasan buatan yang digunakan untuk mempelajari atau memahami sebuah

kata atau bahasa sama seperti manusia berbicara atau menulis dengan bahasa. NLP juga dapat mempelajari perilaku manusia dalam menggunakan bahasa atau mempelajari pola pola terteuntu dalam sebuah data teks. *Natural language processing* digunakan banyak hal dalam era sekarang karena sangat membantu dan bermanfaat bagi manusia seperti penerjamah bahasa, mendeteksi spam, penamaan lokasi, *voice regonition*, *chatbot* atau *bot assistant*. Untuk memahami bahasa memerlukan tahapan karena banyak kosa kata tetapi memiliki arti yang sama atau menggunakan bahasa singkat atau gaul oleh sebab itu *natural language processing* dilakukan ada beberapa tahapan yang dilakukan saat melakukan proses *Natural Language processing yaitu* [17]*:*

2.6.1 Noise Removal

Noise Removal adalah penghapusan *digit* dan simbol atau teks yang dapat menggangu analisis dalam sebuah kalimat. Contoh sebuah kalimat memilki simbol *hastag* atau titik hal ini dapat menjadi masalah karena memiliki *noise* yang mengakibatkan tidak konsisten dalam komputasi atau analisis.

2.6.2 Case Folding

Case Folding adalah proses dimana mengubah huruf besar yang terdapat di kalimat menjadi huruf kecil. Hal ini dilakukan agar setiap kata konsisten. Contoh kalimat “PENGARUH INTERAKSI PERSONAL” dan hasil dari *case folding* adalah “pengaruh interaksi personal” [18].

2.6.3 Tokenization

Tokenization adalah sebuah proses yang bertugas memisahkan sebuah kalimat menjadi sebuah kata atau disebut juga token. *Tokenization* dapat dilakukan berdasarkan pemisah [19]. Tokenization terdiri dari beberapa teknik tokenisasi yaitu:

1. **Word Tokenizer**

Word Tokenizer adalah proses memisahkan kalimat berdasarkan kata [19]. Contoh untuk *word tokenization* sebagai berikut ini, terdapat kalimat "saya mendukung kamu" dan kalimat tersebut dapat dilakukan tokenisasi menjadi ["saya", "mendukung", "kamu"] [18].

2. Sentence Tokenizer

Sentence Tokenizer adalah proses memisahkan kalimat berdasarkan paragraf. Contoh terdapat kalimat "Dia menganalisis perasaan orang. Dia sedang mengerjakan kumpulan data sampel." dan kalimat tersebut dapat dilakukan tokenisasi menjadi ["Dia menganalisis perasaan orang", "Dia sedang mengerjakan kumpulan data sampel"] [19].

2.6.4 Stop Words

Stop words adalah bagian dari filtering yang berisi kata umum yang tidak memiliki makna atau tidak ada efek terhadap komputasi yang signifikan akan dihapus dari suatu kalimat [20]. Contoh dari kata *stop words* untuk berbahasa Indonesia terdapat 3 jenis kata dengan penggunaan terbanyak pada tahun 2016 yaitu kata benda seperti "salon", kata kerja seperti "penyergapan", dan kata sambung seperti "yang" [21].

2.6.5 Stemming

Stemming merupakan proses dari merubah atau menghapus kata akhiran atau imbuhan menjadi kata dasar atau disebut *root word* [20]. Kata akhiran dalam bahasa Indonesia terdapat *Inflection Suffixes* seperti "-lah", "-ku", "-mu" dan *Derivation Suffixes* seperti "-an", "-kan". Contoh proses dari kata *stemming* kata "kesamaan" berubah menjadi "sama", "mencegah" menjadi "cegah", "terjadinya" menjadi "jadi"[22].

2.7 Polarity

Polarity merupakan deteksi polaritas yang paling umum dan penting dari analisis sentimen. *Polarity* menunjukkan rentang angka dari -1 sampai 1, dimana 1 berarti positif, 0 bearti netral, dan -1 berarti negatif. Kalimat subjektif umumnya mengacu pada pendapat pribadi, emosi atau penilaian sedangkan tujuan mengacu pada informasi faktual [23].

2.8 Word Embedding

Word Embedding merupakan kosa kata dokumen yang dapat mengetahui arti dari suatu kata dengan hubungan kata lainnya. Dengan cara mengkonversi sebuah kata menjadi sebuah vektor dari kata tertentu. Salah satu metode yang paling berpengaruh adalah LSI/LSA. *Word Embedding* menghasilkan dimensi vektor yang cukup besar dan terdapat beberapa metode algoritma yang menggunakan *word embedding* seperti *Word2vec* dan *FastText* [24].

2.9 Fast Text

Fast Text merupakan sistem yang cukup populer di kalangan *word embedding* yang dapat mempelajari penyisipan kata dan pengklasifikasi teks. *FastText* mendukung *skip-gram* dengan sampling negatif dan *CBOW*. *FastText* mirip seperti *Word2Vec* karena tujuannya mempelajari representasi vektor dari sebuah kata serta *FastText* merupakan pengembangan dari *Word2Vec* akan tetapi mereka memiliki cara masing-masing untuk memperoleh tujuan. *Word2Vec* menggunakan kata-kata untuk memprediksi kata sedangkan *FastText* memiliki karakter *n-gram* dan penyisipan kata diperoleh dengan menjumlahkan representasi *n-gram* [25].

2.10 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection atau disingkat UMAP merupakan salah satu pembelajaran *manifold* tercepat dan juga merupakan teknik yang digunakan untuk mengurangi dimensi tetapi tidak mengurangi informasi dari

data tersebut. Tujuannya agar dapat dilakukan visualisasi data yang mirip dengan t-SNE secara cepat dan optimal [26].

2.11 Machine learning

Machine learning adalah kecerdasan buatan yang digunakan untuk memecahkan masalah dengan cara mempelajari histori atau data masa lalu dan juga mempelajari pola dari suatu data yang selanjutnya dapat digunakan untuk mengklasifikasikan atau memprediksi suatu kejadian yang berhubungan dengan masalah. *Machine learning* di kelompok menjadi 2 yaitu *supervised learning* dan *unsupervised learning* [27].

Supervised learning merupakan kemampuan algoritma untuk mempelajari data yang tersedia atau memiliki label atau target yang nantinya dapat digunakan untuk memprediksi data baru sedangkan *unsupervised learning* merupakan algoritma pembelajaran yang tidak membutuhkan data berlabel karena metode yang digunakan dengan cara pengelompokan data ke dalam kelompok. Sebagai contoh algoritma yang merupakan *supervised learning* adalah *Naïve bayes*, SVM, K-NN, dll. Untuk algoritma *supervised learning* terdapat algoritma *clustering* seperti K-means [27].

2.12 Support Vector Machine

Support Vector Machine atau SVM merupakan algoritma klasifikasi dan regresi *machine learning* yang bertujuan untuk membuat pembatas antar kelas yang memungkinkan untuk memprediksi label lebih dari satu vektor. Pembatas SVM dapat disebut *hyperplane.* SVM memiliki fungsi kernel yang dapat mempengaruhi kinerja model SVM menjadi baik atau lebih buruk seperti 'poly', 'rbf', 'sigmoid' [28].

2.13 Naïve Bayes

Naïve bayes adalah algoritma *machine learning* yang menggunakan probabilitas sebagai perhitungan untuk mencari kelas tertentu. mendasari secara inheren dalam algoritma pembelajaran adalah bahwa atribut item data independen

satu sama lain. Karena adanya relasi semantik yang kuat di antara kata-kata yang dipilih sebagai fitur, anggapan tersebut justru bertentangan dengan kenyataan, terutama dalam kategorisasi teks. Meskipun demikian, pendekatan tersebut menunjukkan kinerja yang layak dan baik dalam menerapkannya pada kategorisasi teks pada tahun 1997 [15].

2.14 Confusion Matrix

Confusion Matrix berguna untuk memberikan informasi dan evaluasi dari hasil perbandingan klasifikasi yang telah dibuat dengan hasil klasifikasi yang asli. *Confusion Matrix* memiliki 4 nilai yang terdapat *True Positive, False Positive, False Negative*, dan *True Negative* yang dijadikan dalam sebuah tabel mulai dari tabel 2x2 hingga n x n [29].

Class	Y	N
Y	True positive (VP)	False negative (FN)
N	False positive (FP)	True negative (TN)

Class	1	2	3	4	Total
1	70	10	15	5	100
2	8	67	20	5	100
3	0	11	88	1	100
4	4	10	14	72	100

Gambar 2. 2 Confusion Matrix [29]

Pada gambar 2.2 merupakan contoh dari 4 kelas dalam matriks 4x4 yang memiliki keterangan kelas sebagai berikut ini [29]:

1. *True Positive*: terdeteksi sebagai positif dengan benar.
2. *False Positive*: terdeteksi sebagai positif yang seharusnya negatif.
3. *False Negative*: terdeteksi sebagai negatif yang seharusnya positif.
4. *True Negative*: terdeteksi sebagai negatif dengan benar..

Perhitungan *Confusion Matrix* dapat dilakukan agar mengetahui *perfomance metrics* yang terdiri dari *accuracy, precision, recall, dan f1-score.*

2.14.1 Accuracy

Accuracy merupakan alat ukur dari metrik performa yang paling intuitif dan itu hanyalah rasio observasi yang diprediksi dengan benar terhadap jumlah observasi. Akurasi memberi tahu kita seberapa sering kita dapat mengharapkan model pembelajaran mesin kita akan memprediksi hasil dengan benar dari jumlah prediksi itu dibuat. Keakuratan model (melalui kebingungan. Pada gambar 2.2 merupakan perhitungan dengan menggunakan rumus yang diberikan di bawah ini [30].

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \qquad (2.1)$$

2.14.2 Precision

Precision merupakan alat ukur dari metrik performa yang digunakan untuk untuk memprediksi hasil positif dengan benar dari semua prediksi positif yang dibuatnya atau bisa dikatakan presisi menunjukkan seberapa akurat model untuk memprediksi nilai positif. Pada gambar 2.3 merupakan perhitungan dengan menggunakan rumus yang diberikan di bawah ini [30].

$$Precision = \frac{TP}{TP + FP} \qquad (2.2)$$

2.14.3 Recall

Recall merupakan rasio pengamatan positif yang diprediksi untuk memprediksi dengan benar hal-hal positif dari hal-hal positif yang sebenarnya. Recall berguna untuk mengukur kekuatan model untuk memprediksi hasil positif dan itu juga dikenal sebagai sensitivitas model. Kedua tindakan tersebut memberikan informasi yang berharga, tetapi tujuannya adalah untuk meningkatkan daya ingat tanpa mempengaruhi presisi. Pada gambar 2.4 merupakan perhitungan dengan menggunakan rumus yang diberikan di bawah ini [30].

$$Recall = \frac{TP}{TP + FN} \qquad (2.3)$$

2.14.4 F-Measure

F-Measure juga dikenal sebagai nilai-F yang menggunakan skor presisi dan mengingat skor pengklasifikasi. F-measure adalah metrik lain yang umum digunakan dalam pengaturan klasifikasi. F-ukuran dihitung menggunakan harmonik tertimbang berarti antara presisi dan recall. Untuk klasifikasi sikap positif, ada baiknya untuk memahami *tradeoff* antara kebenaran dan cakupan. Rumus umum untuk menghitung F-measure diberikan di bawah ini:

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall} \qquad (2.4)$$

Dalam rumusan di atas, pentingnya setiap istilah dapat diberikan menggunakan nilai yang berbeda untuk b. Nilai yang paling umum digunakan untuk b adalah 1, yaitu dikenal sebagai ukuran F-1. Skor F1 adalah rata-rata tertimbang dari Precision dan Recall. Oleh karena itu, skor ini memperhitungkan positif palsu dan negatif palsu. Secara intuitif tidak semudah memahami akurasi, tetapi F1 biasanya lebih berguna daripada akurasi, terutama jika Anda memiliki distribusi kelas yang tidak merata. Akurasi bekerja paling baik jika positif palsu dan negatif palsu memiliki biaya yang sama. Jika biaya positif palsu dan negatif palsu sangat berbeda, lebih baik melihat *Precision* dan *Recall* [30].

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (2.5)$$

2.15 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
Unwanted Advances In Higher Education: Uncovering Sexual Harassment Experiences In Academia With Text Mining	Information Processing and Management (Vol. 57, 2020) [4]	Amir Karamia, Cynthia Nicole White, Kayla Ford, Suzanne Swan / 2020	Dengan menggunakan metode LDA dan MALLET menunjukkan 45 topik yang sesuai dari 300. Bobot dari nilai topik tersebut berkisar antara 0,0165 hingga 0,0381.	Studi ini mendeteksi dan mengkategorikan topik serta mengeksplorasi variasinya. Teridentifikasi 41 topik dan 5 kategori. Dengan menggunakan metode *teks mining* memudahkan penelitian ini untuk berksplorasi dan memberikan wawasan tentang masalah pelecehan seksual di institusi.
Space Identification Of Sexual Harassment Reports With Text Mining (Vol. 57, 2020) [5]	Proceedings of the Association for Information Science and Technology	Amir Karami, Suzanne Swan, Marcos Moraes / 2020	Hasil dari penelitian ini menunjukkan *Logistic regression* mendapatkan akurasi 83.94% untuk 2 kelas dan 93.26% untuk 7 kelas. Kelas yang dimaskud adalah kelas klasifikasi atau target prediksi.	Penelitian ini untuk mengidentifikasi kategori secara otomatis dari dokumen seksime dengan menggunakan algoritma *multiclass* klasifikasi.

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
A Systematic Literature Review Of Sexual Harassment Studies With Text Mining (Vol. 13, 2021) [31]	Sustainability	Amir Karami, Melek Yildiz Spinel, C. Nicole White / 2021	Dengan menerapkan koherensi analisis untuk menemukan topik dengan LDA, analisis dilakukan pada 2 hingga 50 topik untuk menemukan jumlah topik yang optimal dan menemukan 10 topik tanpa adanya *trend* yang signifikan dan 16 topik ada *trend* yang signifikan salah satunya *sex worker, domestic violence.*	Dari jarak tahun 1977 hingga 2020 yang telah di analisa dan dipelajari sexual harrasment berdasarkan ras, umur, *gender*, dan lokasi.
Analysis Of Sexual Harassment Tweet Sentiment On Twitter In Indonesia Using Naïve Bayes Method Through National Institute Of Standard And Technolog	Journal of Advances in Information Systems and Technology	Kholiq Budiman, Nurul Zaatsiyah, Ulfatun Niswah / 2020	Dengan menggunakan metode algoritma *naïve bayes* mendapatkan hasil akurasi 83% dan presisi 57% dari 210 data training dan 90 data testing dimana total data adalah 300. Dari hasil tersebut 31,3% memilki sentimen positive dan	Twitter adalah tempat sosial media yang dimana dapat berkomunikasi secara bebas melewat tweet serta bebas untuk melakukan pelecehan seksual melalui *tweet* atau dm. dari hasil sentimen analisis dengan menggunakan metode *naïve bayes* 69.7%

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
y Digital Forensic Acquisition Approach (Vol .2, 2020) [32]			69,7% sentimen negatif.	merupakan sentimen negatif dengan akurasi 83%. Dari hasil tersebut dapat menunjukan twitter masih belum bijak dalam memberikan sanksi kepada pengguna twitter yang memberikan informasi negatif .
Evaluating Frameworks For Implementing Machine Learning In Signal Processing: A Comparative Study Of CRISP-DM, SEMMA And KDD (2018) [9]	EXAMENSARBETE INOM TEKNIK	Antonia Dåderman, Sara Rosander / 2018	Hasil dari komparasi ketiga *framework* yang dimana CRISP-DM lebih cocok digunakan dibidang IT, *Medicine*, *Software*.untuk SEMMA hanya bagus di bidang IT dan *Customer Care* sedangkan KDD cocok di industri transportasi dan turis. CRISP-DM juga memiliki banyak kelebihan seperti proses yang jelas, cocok	Hal ini menghasilkan kesimpulan bahwa CRISP-DM adalah kerangka kerja yang paling cocok untuk Saab karena berasal dari perspektif bisnis, merupakan metode iteratif dan mudah diimplementasikan ke dalam proses pengembangan yang digunakan Saab saat ini. CRISP-DM juga terstruktur dengan baik dengan langkah-

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
			penggunaan *data mining*, memiliki dokumentasi yang baik sedangkan SEMMA memiliki *iterative process* dan mendukung *data mining*.	langkah yang terdefinisi dengan baik. Jajak pendapat menunjukkan bahwa CRISP-DM adalah salah satu kerangka kerja yang paling banyak digunakan, yang menurut kami memperkuat kesimpulan kami.
Comparation Of Classification Algorithm On Sentiment Analysis Of Online Learning Reviews And Distance Education (Vol. 18, 2021) [8]	Jurnal TECHNO Nusa Mandiri	Lila Dini Utami1, Siti Masripah Sistem / 2021	Dengan menggunakan *dataset* sebesar 300 yang dibagi menjadi 2 yang dimana terdapat 150 sentimen positif dan 150 sentimen negatif. Hasil prediksi yang dilakukan berdasarkan ketiga algoritma yang dimana SVM mendapatkan akurasi sebesar 87,67%, K-NN mendapatkan akurasi sebesar 86, 33% dan *Naïve Bayes* mendapatkan 83,33%. Dapat dikatakan	Berdasarkan penelitian kami, dengan membandingkan tiga algoritma klasifikasi yaitu algoritma *Naïve Bayes* (NB), Algoritma *k-Nearest* Neighbor (k-NN) dan algoritma *Support Vector Machine* (SVM), dapat diketahui bahwa hasil dari Algoritma *Naïve Bayes* (NB) mendapatkan Nilai akurasi: 83,33% +/- 4,44% (mikro

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
			bahwa dalam penelitian ini SVM mendapatkan hasil yang lebih baik.	rata-rata: 83,33%), dengan nilai AUC 0,756. Sedangkan algoritma *k-Nearest Neighbor* (k-NN) mendapatkan nilai *Accuracy*: 86,33% +/- 6,93% (micro *average*: 86,33%) dengan nilai AUC 0,911. dan untuk Algoritma *Support Vector Machine* (SVM) diperoleh nilai *Accuracy*: 87,67% +/- 6,49% (*micro average*: 87,67%) dengan nilai AUC 0,939.
Sentiment Analysis In The Sales Review Of Indonesian Marketplace By Utilizing Support Vector Machine (Vol. 4, 2018) [33]	Journal of Information Systems Engineering and Business Intelligence	Lutfi, Anang Anggono Permanasari, Adhistya Erna Fauziati, Silmi / 2018	Dalam penelitian ini hasil yang didapatkan dari komparasi kedua algoritma *naive bayes* dan SVM dengan menggunakan *dataset* berbahasa Indonesia mendapatkan	Ulasan penjualan sering ditulis dalam teks yang tidak terstruktur. Analisis sentimen dapat digunakan untuk mengetahui polaritas teks-teks yang tidak

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
			akurasi dengn rata-rata 93% untuk SVM sedangkan *naïve bayes* 90%.	terstruktur. *Review* penjualan dapat diambil dari berbagai sumber seperti media sosial. Pada penelitian dengan domain yang sama namun dengan *dataset* dan metode yang berbeda dilakukan oleh Fiarni et.al [3]. Kumpulan data ulasan toko online yang dikumpulkan dari media sosial. Metode yang digunakan adalah *Naive Bayes* Classifier. Akurasi penelitian ini adalah 89,21%. Namun dalam penelitian ini, data diambil dari situs *marketplace*. Ini akan menggambarkan penjualan yang sejalan dengan lebih baik. Ilustrasi penjualan yang lebih baik akan memberikan

				gambaran penjualan yang lebih jelas. Hasilnya sesuai dengan pernyataan peneliti [2] dan [14] yang menunjukkan bahwa SVM dan NB memberikan hasil yang baik dalam klasifikasi teks. Perbandingan antara SVM dan NB menunjukkan bahwa rata-rata akurasi SVM lebih tinggi dari NB. Rata-rata mencapai 93,65% menggunakan 25% fitur dengan TF-IDF tertinggi. Hal ini menunjukkan bahwa ekstraksi fitur penting dapat mempengaruhi hasil akurasi. Hasil ini menegaskan hasil penelitian [15] yang menunjukkan bahwa SVM dengan *kernel* linier memberikan akurasi yang

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
				lebih tinggi dibandingkan dengan NB. NB menunjukkan bahwa peningkatan akurasi sejalan dengan peningkatan jumlah fitur yang digunakan dalam analisis. Sedangkan pada SVM, peningkatan jumlah fitur belum tentu menunjukkan peningkatan akurasi. VI.
Analisa Akurasi Permodelan Supervised Dan Unsupervised Learning Menggunakan Data Mining (Vol. 23, 2017) [34]	Journal of Computer Science and Computer Engineering STMIK Widya Cipta Dharma	Warnia Nengsih / 2017	membandingkan hasil akurasi dari *supervised learning* yang menggunakan algoritma regresi linear, *decision tree*, dan SVM mendapatkan akurasi dengan rata-rata sebesar 82,33% sedangkan *unsupervised learning* dengan algoritma *K-Means*, *single linkage*, dan *apriori*	Dengan menggunakan bahasa pemograman matlab untuk membandingkan kedua teknik *machine learning*, teknik yang terbaik adalah teknik *supervised learning* dengan rata-rata akurasi sebesar 82,33% sedangkan *unsupervised* memperoleh

Judul Jurnal	Nama Jurnal	Penulis / Tahun	Hasil	Kesimpulan
			mendapatkan hasil akurasi dengan rata-rata sebesar 78%.	rata-rata akurasi 78%. Nilai akurasi juga dipengaruhi oleh keberagaman jenis data.
Perbandingan Kinerja Word Embedding Word2vec, Glove, Dan Fasttext Pada Klasifikasi Teks (Vol 14, 2020) [10]	Jurnal TEKNOKOMPAK	Arliyanti Nurdin, Bernadus Anggo Seno Aji, Anugrayani Bustamin, Zaenal Abidin / 2020	Hasil dari eksperimen, kinerja CNN dalam mengklasifikasikan teks menggunakan *word embedding word2vec, GloVe, dan FastText* menggunakan ukuran *F-Measure* secara berturut-turut untuk *dataset* 20 *newsgroup* adalah 0.925, 0.958, dan 0.979, dan *dataset Reuters. News* adalah 0.694, 0.688, dan 0.715.	Dengan menggunakan algoritma CNN untuk mengklasifikasi teks. Dari ketiga algoritma *word embedding* yang memiliki kinerja yang baik adalah *FastText* yang lebih unggul dalam *word embedding* yang telah di uji coba dengan 2 *dataset* berbeda.

Berdasarkan penelitian terdahulu yang terdapat pada tabel 2.1 yang dimana merupakan referensi untuk melakukan penelitian ini. Artikel jurnal dengan judul *Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A Comparative Study of CRISP-DM, SEMMA and KDD* yang akan menjadi referensi untuk memilih *framework data mining* [9] dan artikel jurnal dengan berjudul *Applying CRISP-DM Process Model A Systematic Literature Review on*

Applying CRISP-DM Process Model akan digunakan sebagai implementasi *framework data mining* yang sesuai dengan standar[13].

Pada penelitian terdahulu telah melakukan eksperimen komparasi dari berbagai algoritma *word embedding* yaitu *Word2Vec, Glove,* dan *FastText* dengan menggunakan 2 datasets berbeda (20 *newsgroups* dan *Routers*). Dari hasil penelitian terdahulu kinerja *FastText* lebih unggul ditimbang Glove dan Word2Vec karena hasil dari *F1-score* yang didapatkan oleh *FastText* adalah 0.979 untuk *datasets* 20 *newsgroup* sedangkan untuk *datasets routers* mendapatkan 0.715 [10].

Teknik *machine learning* yang digunakan pada penelitian ini adalah *supervised learning* yang berdasarkan penelitian terdahulu dengan judul Analisa Akurasi Permodelan *Supervised* Dan *Unsupervised Learning* Menggunakan Data Mining yang dimana *supervised learning* lebih baik 4,33% dari *unsupervised learning*.

Selanjutnya, Referensi yang digunakan untuk dijadikan acuan pemilihan algoritma *machine learning* dengan judul *Comparation Of Classification Algorithm On Sentiment Analysis Of Online Learning Reviews And Distance Education* [8] dan *Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine* Sebagai pendukung kedua membuktikan bahwa algoritma SVM lebih baik ditimbang *Naïve bayes* dalam topik sentimen analisis berbahasa Indonesia [8]. Dengan jurnal tambahan yang berujudul *Analysis of Sexual Harassment Tweet Sentiment on Twitter in Indonesia using Naïve Bayes Method through National Institute of Standard and Technology Digital Forensic Acquisition Approach* yang membahas sentimen analisis berbahasa Indonesia dengan algoritma *naïve bayes* mendapatkan akurasi yang mirip yaitu sebesar 83% [33].