BAB II

PENELITIAN TERDAHULU

2.1. *Related Theories*

2.1.1 *Text Mining*

Text mining adalah proses intensif dimana *user* dapat berinteraksi dengan dokumen – dokumen dan menggunakan sarana analisis untuk menemukan pola dari minat. Pada *Text Mining* biasanya peneliti melakukan pendekatan untuk mencari pengetahuan didalam teks berjumlah besar [8]. *Text mining process* berunsurkan beberapa langkah [8] :

- *Defining the problem*

  Pada tahap ini masalah pada domain perlu dimengerti dan pertanyaan yang adaperlu dijawab atau didefinisikan.

- *Collecting the necessary data*

  Data yang dikumpulkan harus sesuai yang diinginkan dan sudah di identifikasi. *Data* juga perlu di dokumenasi agar siap untuk di analisa lebih lanjut.

- *Defining features*

  Pendefinisian akan teks yang sudah mengkarakterisasikan fungsi dari teks itusendiri.

- *Analyzing the data*

  Tahapan dimana *data* di proses agar dapat ditemukan pola yang ada. Untukmengatasi suatu masalah, biasa diperlukan beberapa model. Setiap model

memiliki karakteristik yang berbeda. Model yang terpakai dapat bersifat *whitebox* dan *black box*. Model yang cocok biasanya akan bergantung secara kuat dengan *data*.

- *Interpreting the result*

Hasil dapat di temukan dari analisis. Pada tahap ini diperlukan adanya verifikasi dan validasi untuk meningkatkan *reability of results*.

2.1.2 *Sentiment Analysis*

Analisis Sentimen adalah *Natural Languange Processing* (NLP) yang membangun sistem untuk mengenali dan mengekstraksi opini dalam bentuk teks. Informasi berbentuk teks saat ini banyak terdapat di *internet* dalam *format forum*, *blog*, *media sosial*, serta situs berisi *review*. Dengan bantuan *sentiment analysis*, informasi yang tadinya tidak terstruktur dapat diubah menjadi *data* yang lebih terstruktur.

Data tersebut dapat menjelaskan opini masyarakat mengenai produk, merek, layanan, politik, atau topik lainnya. Perusahaan, pemerintah, maupun bidang lainnya kemudian memanfaatkan *data-data* tersebut untuk membuat analisis *marketing*, *review* produk, umpan-balik produk, dan layanan masyarakat[9].

2.1.3 *Naïve Bayes Classifiers*

Menurut teori, pengklasifikasi Bayes yang optimal memberikan hasil klasifikasi terbaik berdasarkan *data* dan hipotesis yang diberikan. Hal

ini terutama (tetapi tidak hanya) benar dalam penambangan teks karena menetapkan *rect class-label* c j bergantung pada jumlah atribut, yaitu kata-kata dalam kontribusi media sosial atau dokumen teks. Setelah mengumpulkan sejumlah besar sampel pelatihan (yang bagus dari titik komputasi probabilitas *vie*), kosakata mereka mungkin sangat besar - ribuan atau puluhan ribu kata unik (atau istilah umum). Setiap kata memiliki probabilitasnya di setiap kelas dan kemungkinannya sangat banyak kombinasi atribut, meskipun *data*nya dikurangi dengan, misalnya, menghilangkan suku-suku yang tidak penting. Rumus dapat dilihat pada rumus 2.1 :

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

Rumus 2. 1 Rumus *Naive Bayes* 1

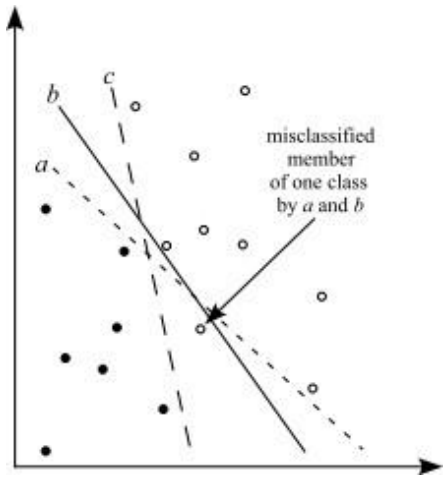
Untuk nilai N yang besar, kompleksitas komputasi diberikan oleh *item* diatas. Karena kemungkinan saling ketergantungan kondisional atribut maka perlu untuk menghitung probabilitas dari semua kemungkinan kombinasi mereka untuk *c j* tertentu. *IDE* yang disebut pengklasifikasi *na¨ive Bayes*, yang (secara teoritis tidak cukup akurat) berdasarkan asumsi bahwa tidak ada interdependensi antar atribut (yaitu *'naïvitas*') yang dihasilkan sebagai cara untuk membuat komputasi menjadi lebih mudah dan dapat diterapkan secara praktis. Artinya, untuk klasifikasi bisa saja untuk menggunakan persamaan yang lebih sederhana untuk menentukan cMAP, yang dapat disebut sebagai *Naïve Bayes* seperti pada rumus 2.2 [8] :

$$c_{NB} = \underset{c_j \in C}{argmax}\, p(c_j) \prod_i p(a_i|c_j) \,.$$

Rumus 2. 2 Rumus *Naive Bayes* 2

2.1.4 *Support Vector Machine*

Support Vector Machines, SVM, adalah algoritma sukses lainnya yang telah membuktikan dan mendemonstrasikan hasil luar biasa di banyak aplikasi dunia nyata yang berbeda. Dasarnya berasal dari zaman yang relatif awal, dimulai pada tahun 1963, ketika matematikawan Rusia (Soviet) Vladimir Vapnik (bersama dengan Alexey Chervonenkis, Alexander Lerner, dan lainnya) menerbitkan karya tentang pembelajaran statistik dan pengenalan pola. Karya-karya ini secara bertahap memunculkan salah satu algoritme yang berhasil saat ini berdasarkan teori aljabar linier. Pada prinsipnya, *SVM* adalah algoritma klasifikasi yang bertujuan untuk memisahkan dua kelas menggunakan *hyperplane* - pengklasifikasi linier. Ada beberapa cara untuk menempatkan *hyperplane* (atau garis lurus dalam ruang dua dimensi) yang memisahkan elemen dari dua kelas. Masalahnya adalah jika pilihan seperti itu ada, seseorang secara umum dapat membuat



Gambar 2. 1 Diagram SVM

banyak batasan seperti itu untuk serangkaian contoh pelatihan yang diberikan.

Seperti pada gambar 2.1, menurut *IDE SVM*, perlu dicari zona terluas (sabuk) memisahkan kedua kelas. Contoh-contoh yang dipilih secara tepat yang disebut vektor pendukung menentukan tepi zona ini [8].

2.1.5 *Data Pre-processing*

Merupakan tahap yang harus dilakukan agar *data* yang kita dapat siap diolah. Pada tahap ini biasanya terdapat 5 *Sub-Process* yaitu *Data cleansing* dimana membuat *data* yang tidak relevan menjadi relevan, *Case folding* membuat huruf menjadi bentuk yang sama, *tokenization* memisahkan kalimat menjadi *token*, *stop words removal*, dan terakhir *stemming* dimana proses menghilangkan *suffix* dan *affix*[10].

2.1.6 *Rapid Miner*

Merupakan aplikasi yang memiliki fungsi untuk *Machine learning*, preparasi *data*, *text mining*, dan analisi prediktif. Aplikasi ini biasa digunakan untuk tujuan yang berbeda seperti pelatihan, pendidikan, bisnis dan komersial dan lain – lain[11].

2.1.7 *K-Fold Cross-Validation*

Validasi silang 10 kali lipat adalah standar untuk menguji akurasi pengklasifikasi secara bertahap. Ada 10 langkah, masing-masing menggunakan 10% sampel untuk pengujian dan 90% untuk pelatihan, dan setiap sampel secara bertahap mengambil bagian di kedua bagian. Metrik kinerja klasifikasi dihitung disetiap langkah. Nilai akhirnya diberikan oleh nilai rata-rata dari semua 10 pengukuran. Dalam contoh kami, hanya nilai akurasi yang dipertimbangkan untuk kesederhanaan.

Kesalahan standar *mean* (*SEM*) yang digunakan dalam statistik untuk menyatakan kesalahan rata-rata untuk sampel pengujian yang dipilih secara acak dari populasi *data* (generalisasi kesalahan untuk seluruh populasi) juga dihitung. Diberikan di bawah ini adalah contoh *R-code*. Seorang pembaca akan melihat bahwa kode untuk pengujian validasi silang sangat mirip dengan kode yang menggunakan set pengujian untuk pengujian[8].

2.1.8 *Confusion Matrix*

Confusion Matrix sebuah cara perhitungan yang sering dipakai pada *machine learning* untuk mengklasifikasi tingkah laku dari *classification models*. *Confusion Matrix* biasanya di diterapkan untuk mengevaluasi performa dari klasifikasi pada *datasets*. *Confusion Matrix.* Ada juga peneliti yang menggunakan *Confusion Matrix* untuk membedakan nilai yang diprediksi dan nilai asli dari *model element* pada *software engineering*. Pada *Confusion Matrix* terdapat 4 cara pengukuran yaitu *TP*, *FP*, *TN*, dan *FN* yang dipakai pada *program - program Java*[12].

2.1.9 *TF-IDF*

TF-IDF merupakan algoritma yang secara luas dipakai untuk pemilihan fitur dalam *text information processing*. *TF* merupakan singkatan dari *Term Frequency* yang merepresentasikan seberapa sering munculnya istilah – istilah fitur dari kumpulan teks, dan *IDF* merupakan singkatan dari *Inversed Document Frequency* yang merupakan pengukur dari kepentingan umum suatu istilah yang sering muncul di dalam *data set*[13].

2.1.10 *Instagram*

Salah satu media sosial yang sering dipakai masyarakat adalah *instagram*, dimana pengguna dapat berbagi foto dan *video* secara gratis dengan pengguna *instagram* lainnya. Pengguna juga dapat melihat , mengomentari, dan menyukai foto yang di berikan pengguna lainnya. Pada penelitian ini *comment* pada *instagram* akan di pakai untuk mengambil *data* dengan menggunakan *instagram API*. Dengan *API* tersebut memungkinkan untuk mengambil *data data* komen pada *instagram*[14].

2.1.11 *Tableau*

Tableau merupakan aplikasi yang dipakai untuk menyajikan suatu *data* dengan tampilan yang lebih menarik. Dengan *Tableau user* dapat memakai *data source* secara *virtual* dan dapat mengolah *data* dalam berbagai bentuk, besar, dan tipe [15].

2.1.12 *Pycharm*

Pycharm merupakan aplikasi populer yang biasa digunakan untuk menulis *script* berbahasa *phyton*. Pada aplikasi ini juga memungkinkan untuk *import libraries* seperti *pandas, numpy, sklearn, tinker, mtlab plot*. *User* hanya perlu menjalani kode yang di buat untuk melihat hasil kode tersebut[16].

2.2. *Previous Research*

Tabel 2. 1 Penelitian Terdahulu

1	**Penulis**	Sheeba Naz, Aditi Sharan, Nidhi Malik
	Nama Jurnal	*2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI) sentiment*

	Judul	*sentiment classification on Twitter data using Support Vector Machine*
	Metode	*- SVM*
	Kesimpulan	*Purposed approach* memiliki akurasi 81% lebih tinggi dari pendekatan lain.
2	**Penulis**	Veny Amilia Fitri, Rachmadita Andrewari, Muhammad Azani Hasibuan.
	Nama Jurnal	*The Fifth Information Systems International Conference 2019.*
	Judul	*Sentiment Analysis of Social Media Twitter with Case of Anti- LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm*
	Metode	*- Naïve Bayes* *- Descision Tree* *- Random Forest*
	Kesimpulan	Komentar mengenai *Anti-LGBT* cenderung termasuk dalam kategori netral.
3	**Penulis**	Pratiwi, Nur Indah Budi, Indra Alfina, Ika
	Nama Jurnal	*2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*
	Judul	*Hate speech detection on Indonesian instagram comments using FastText approach*
	Metode	*- FastText Approach* *- SVM*
	Kesimpulan	*Data set* yang baru mengenai *Hate speech detection* pada komen *instagram*.
4	**Penulis**	Alrumaih, Abdulrahman Al-Sabbagh, Ali Alsabah, Ruaa Kharrufa, Harith Baldwin, James
	Nama Jurnal	*International Journal of Electrical and Computer Engineering*
	Judul	*Sentiment analysis of comments in social media*
	Metode	*- Sentiment Analysis* *- Data Processing*
	Kesimpulan	Arabic *comment* data akan lebih akurat jika di ukur bersama emoji dan teks.

5	**Penulis**	Joviano Siahaan, Wella, Ririn I. Desanti
	Nama Jurnal	*ULTIMA InfoSys, Vol. XI, No. 2 |*
	Judul	Apakah Youtuber Indonesia Kena Bully Netizen?
	Metode	- *SVM*
	Kesimpulan	*Youtuber* Indonesia tidak melulu dirundung oleh para masyarakat Indonesia.

Pada tabel 2.1 terdapat 4 jurnal terdahulu yang dipakai sebagai acuan untuk berjalannya penelitian ini. Pada jurnal nomor 1 yang ditulis oleh Sheeba Naz akan mengaplikasikan penggunaan *SVM*. Pada jurnal nomor 2 yang ditulis oleh Veny Amilia Fitri akan diambil pengaplikasian *Naive bayes*. Pada jurnal nomor 3 yang ditulis oleh Nur Indah Budi akan diambil penelitian pada *Instagram* dan penggunaan *SVM*. Pada jurnal nomor 4 yang ditulis oleh Alrumaih akan diambil *sentiment analysis* dan *data processing*. Pada jurnal nomor 5 yang ditulis oleh Joviano akan diambil bagian pemakaian *SVM* dan penggunaan *social media* sebagai objek analisa.

Dari 5 jurnal terdahulu terdapat perbedaan dari penelitian yang akan dilakukan. Dalam penelitian ini akan dilakukan *sentiment analysis* dengan menggunakan 2 algoritma yaitu *naive bayes* dan *SVM* pada *comment instagram* pada akun *Provider Internet* Indonesia sebagai objek yang diteliti. Dari masing – masing algoritma akan dibandingkan perbedaan hasil serta akurasi yang didapatkan.