BAB II

LANDASAN TEORI

2.1 Analisis Sentimen

Analisis sentimen adalah penggunaan pemprosesan bahasa alami, analisis teks untuk mengidentifikasi, mengumpulkan, mengekstrak, menghitung, serta mempelajari suatu data untuk menjadi sebuah informasi subjektif [15].

Analisis sentimen terbagi menjadi 3 level [16] yaitu:

1. *Level* Dokumen

   Menganalisis dan mengklasifikasikan dokumen tersebut memiliki sentimen positif atau negatif. Level ini sangat cocok diterapkan untuk membandingkan lebih dari satu entitas.

2. *Level* Kalimat

   Menganalisis dan menentukan suatu kalimat apakah kalimat tersebut bernilai sentimen positif, netral, atau negatif. Kalimat yang bernilai netral berarti kalimat tersebut bukan merupakan sebuah opini.

3. *Level* Entitas dan Aspek

   Pada level ini tidak dilakukannya proses analisis tetapi level ini bertujuan untuk menentukan sentimen entitas pada tiap aspek yang sedang dibahas.

2.2 *Support Vector Machine* (SVM)

SVM merupakan suatu teknik yang baru untuk melakukan sebuah prediksi dalam kasus klasifikasi ataupun regresi, model ini masuk kedalam kelas *supervised learning* yaitu pada implementasinya diperlukan sebuah tahap pelatihan menggunakan *sequential training* SVM dan baru dilanjutkan dengan proses pengujian [6].

Berikut adalah merupakan kelebihan dari metode SVM [17] yaitu:

1. Memiliki kemampuan generalisasi yang cukup tinggi.
2. Dengan metode ini mampu menghasilkan sebuah klasifikasi yang baik meskipun harus dilatih dengan data yang sedikit dengan parameter yang sederhana. SVM juga memiliki konsep dan formulasi yang jelas.
3. Metode ini mudah untuk diimplementasikan karena dapat dirumuskan ke dalam masalah QP (*Quadratic Programming).*

2.3 Python

Python merupakan sebuah bahasa pemrograman yang memiliki banyak manfaat dalam mendukung pemrograman yang berorientasi objek dan dapat berjalan diberbagai macam *platform* sistem operasi seperti PCs, Macintosh, UNIX, dan lainnya [18].

Python memiliki *natural language toolkit (NLTK)* yang digunakan untuk pemrosesan dan klasifikasi teks serta *Python* juga menyediakan *library* yang dapat digunakan dalam membantu klasifikasi data, pembelajaran mesin (*machine learning)* dan juga *data mining* [19].

Bahasa pemrograman *python* ini memiliki beberapa kelebihan yaitu [18]:

1. Pengembangan program yang dilakukan terjadi dengan begitu cepat dan coding yang dibutuhkan lebih sedikit dibandingkan dengan bahasa pemrograman yang lain.
2. Mendukung *multiplatform* sistem operasi.
3. Memiliki sistem pengelolaan terhadap penggunaan memori secara otomatis.
4. Bahasa pemrograman *python* ini bersifat *Object Oriental Programming (OOD).*

2.4 *Term Frequency - Inverse Document Frequency* (TF-IDF)

TD-IDF merupakan suatu metode yang digunakan untuk melakukan pembobotan pada setiap kata pada dokumen dan metode ini juga sering digunakan karena dinilai mudah dan memiliki hasil yang cukup akurat [20].

TF (*Term-Frequency)* lebih memfokuskan pada setiap kata yang sering muncul dalam dokumen. Sedangkan IDF (*Inverse Document Frequency)* ini memfokuskan dalam memberikan bobot yang rendah untuk kata yang muncul dalam dokumen tersebut [21].

2.5 N-Gram

N-Gram merupakan sebuah metode yang digunakan untuk menampung potongan kata-kata dari kalimat sesuai dengan jumlah karakternya [22]. N-Gram berfungsi dengan baik dan tidak terlalu sensitif walaupun terdapat adanya kesalahan pada tulisan atau kalimat yang ada, N-Gram juga diyakini dapat berjalan dengan efisien dan berproses dengan cepat [22].

Dalam metode N-Gram terdapat beberapa jenis yaitu *unigram* dengan n=1, *bigram* dengan n=2, dan *trigram* dengan n=3 dan penerapan N-Gram ini dalam melakukan klasifikasi diduga dapat meningkatkan hasil kinerja klasifikasi [23].

2.6 Text / Data Preprocessing

Data Preprocessing adalah sebuah teknik *data mining* yang bertujuan untuk mengubah atau mengolah data yang masih mentah menjadi sebuah format data yang mudah untuk dimengerti, teknik ini diperlukan untuk menyelesaikan jenis masalah pada data, seperti *noisy data*, data yang berulang atau sama, serta nilai data yang hilang [21].

Adapun beberapa tahapan yang dilakukan dalam proses *Data Preprocessing* [24] antara lain:

1. Formalisasi dan Translasi

   Formalisasi merupakan tahap mengubah kata menjadi bentuk kata baku yang sesuai dengan KBBI. Sedangkan translasi adalah tahap untuk menerjemahkan kata dari bahasa asing ke bahasa Indonesia.

2. *Cleansing*

   Tahap ini bertujuan untuk menghilangkan elemen yang tidak diperlukan pada data yang nantinya akan di proses analisis sentimen.

3. *Tokenizing*

   Tahap ini bertujuan untuk memisahkan teks atau kalimat ulasan menjadi sebuah potongan-potongan kata.

4. *Case Folding*

   Case Folding bertujuan untuk mengubah kata-kata yang dihasilkan pada tahap *tokenizing* menjadi sebuah karakter huruf kecil.

5. *Stemming*

   Stemming bertujuan untuk menghilangkan kata imbuhan pada data yang telah dikumpulkan agar menjadi sebuah kata dasarnya.

6. *Stopword Removal*

   Tahap ini bertujuan untuk menghilangkan kata-kata yang tidak terlalu berpengaruh pada analisis sentimen atau kata yang terdapat di dalam *stopword list.*

2.7 Scraping

Scraping merupakan suatu metode yang dapat digunakan untuk mengumpulkan dan mengekstrak data atau informasi dari situs yang dilakukan secara otomatis [25].

Tujuan digunakannya *scraping* ini adalah untuk menggali semua informasi yang ada dari situs yang berbeda dan yang tidak terstruktur lalu akan ditransformasikan ke dalam bentuk data yang lebih rapi dan terstruktur ke dalam format *database, spreadsheets,* dan *comma separated values (CSV)* [25].

Scraping merupakan sebuah teknik untuk menggali sebuah informasi dari suatu situs yang ada, *web scraping* ini mengambil informasi dengan cara menelusuri dokumen-dokumen *HTML* yang ada di situs tersebut yang nantinya akan diambil untuk dijadikan sebuah informasi dan di *tag* ke *HTML* agar bisa

mendapatkan informasi untuk ditirukan ke dalam aplikasi *scraping* yang akan dibuat [26].

2.8 Confusion Matrix

Confusion Matrix merupakan suatu metode yang sering digunakan pada penelitian untuk melakukan proses perhitungan tingkat akurasi pada *data mining* [21]. *Confusion Matrix* adalah sebuah metode yang dapat digunakan untuk melakukan proses pehitungan akurasi dan selain akurasi metode ini dapat juga digunakan untuk mencari nilai *recall, precision,* dan *error rate* [20].

Confusion Matrix adalah suatu metode untuk melakukan proses perhitungan akurasi dan juga dapat digunakan untuk mengukur serta mengevaluasi performa kinerja dari klasifikasi [27].

Berikut adalah Tabel 2.1 yang merupakan rumus dari *confusion matrix* [20] antara lain;

Tabel 2. 1 Rumus *Confusion Matrix*

Accuracy	$\frac{(\boldsymbol{TP}+\boldsymbol{TN})}{(\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{TN}+\boldsymbol{FN})}$
Recall	$\frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}}$
Precision	$\frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}}$
F1_Score	$\frac{(\boldsymbol{TP})}{\boldsymbol{TP}+\mathbf{1/2}(\boldsymbol{FP}+\boldsymbol{FN})}$

Berikut merupakan penjelasan dari rumus yang digunakan pada tabel 2.1 adalah sebagai berikut:

- Akurasi : $\frac{(TP+TN)}{(TP+FP+TN+FN)}$

  Pencarian nilai akurasi dilakukan untuk menampilkan rasio dari jumlah prediksi data yang diklasifikasi dengan benar dari data yang digunakan [23].

- *Recall* : $\frac{TP}{TP+FN}$

  Recall digunakan untuk mencari nilai positif yang berhasil diprediksi dari total nilai positif aktual [28].

- *Precision* : $\frac{TP}{TP+FP}$

  Precision digunakan untuk mencari nilai positif dari hasil prediksi data yang memberikannya nilai positif [28].

- *F1_Score* : $\frac{(TP)}{TP+1/2(FP+FN)}$

  F1_Score disini digunakan untuk menghitung hasil rata-rata dari hasil *precision* dan *recall* [28].

Keterangan dari istilah pada tabel 2.1 rumus *confusion matrix* [29]:

TP (*True Positive*) : Jumlah data yang sebenarnya bernilai positif dan hasil prediksi juga bernilai positif.

TN (*True Negative*) : Jumlah data yang sebenarnya bernilai negatif dan hasil prediksi juga bernilai negatif.

FP (*False Positive*) : Jumlah data yang sebenarnya bernilai negatif dan hasil prediksinya bernilai positif.

FN (*False Negative*) : Jumlah data yang sebenarnya bernilai positif dan hasil prediksinya bernilai negatif.

2.9 Penelitian Terdahulu

Pada penelitian kali ini, menggunakan beberapa penelitian terdahulu atau penelitian yang sudah dilakukan sebelumnya untuk memperkuat proses penelitian yang sedang dilakukan, dan penelitian terdahulu ini sebagai acuan untuk mendapatkan teori-teori yang membantu dalam melakukan penelitian kali ini. Berikut adalah Tabel 2.2 hingga Tabel 2.7 yang akan menyajikan beberapa penelitian terdahulu yang memiliki keterkaitan dan yang akan digunakan pada penelitian kali ini.

Tabel 2. 2 Penelitian Terdahulu 1

Penelitian 1	
Judul	ANALISIS SENTIMEN OPINI PUBLIK BAHASA INDONESIA TERHADAP WISATA TMII MENGGUNAKAN NAÏVE BAYES DAN PSO
Peneliti & Tahun Penelitian	Ratih Yulia Hayuningtyas, Retno Sari (STMIK Nusa Mandiri Jakarta), Tahun 2019.
Sumber	Jurnal TECHNO Nusa Mandiri Vol.16, No.1
Tujuan	Untuk menganalisis sentimen terhadap ulasan atau opini publik tentang objek wisata TMII yang nantinya akan dijadikan sebuah informasi yang memudahkan para wisatawan dalam memilih objek wisata yang ingin dikunjungi.
Metode	Metode klasifikasi *Naïve Bayes* dan *Naïve Bayes* dengan PSO (*Partcicle Swarm Optimization*)
Hasil Penelitian	Mendapatkan sebuah perbandingan 2 akurasi dari metode yang digunakan. Dari metode *Naïve Bayes* tanpa PSO didapatkan akurasi sebesar 70% dengan menggunakan 4 *Fold Cross Validation*. Sedangkan metode *Naïve Bayes* dengan PSO didapatkan akurasi sebesar 94.02% dengan menggunakan 9 *Fold Cross Validation*.

Tabel 2. 3 Penelitian Terdahulu 2

Penelitian 2	
Judul	ANALISIS SENTIMEN BERBASIS ASPEK ULASAN PELANGGAN TERHADAP KERTANEGARA PREMIUM GUEST HOUSE MENGGUNAKAN SUPPORT VECTOR MACHINE
Peneliti & Tahun Penelitian	Wirdhayanti Paulina, Fitra Abdurrachman Bachtiar, Alfi Nur Rusydi (Universitas Brawijaya), Tahun 2020.
Sumber	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol.4, No.4
Tujuan	Untuk mengetahui opini publik terhadap Kertanegara Premium Guest House bernilai positif atau negatif sehingga nantinya dapat dijadikan sebuah acuan untuk evaluasi tentang pemberian layanan, akomodasi, dan fasilitas yang baik kepada para customernya.
Metode	Metode klasifikasi *Support Vector Machine (SVM)*
Hasil Penelitian	Dari hasil pengujian menggunakan metode *Support Vector Machine* (SVM) didapatkan metric *F1-Score*, *Precision*, *Recall*, serta Akurasi diatas 70%.

Tabel 2. 4 Penelitian Terdahulu 3

Penelitian 3	
Judul	KLASIFIKASI SENTIMEN WISATAWAN CANDI BOROBUDUR PADA SITUS TRIPADVISOR *MENGGUNAKAN SUPPORT VECTOR MACHINE* DAN *K-NEAREST NEIGHBOR*
Peneliti & Tahun Penelitian	Rahayu Prihatini Saputri, Wiwiek Setya Winahju, dan Kartika Fithriasari (Departemen Statistika, Fakultas Matematika Komputasi dan Sains Data, Institut Teknologi Sepuluh Nopember (ITS)) Tahun 2019
Sumber	JURNAL SAINS DAN SENI ITS Vol. 8, No. 2
Tujuan	Ingin mengetahui persepsi para wisatawan dari opini yang ada terkait dengan Candi Borobudur.
Metode	Metode klasifikasi *Support Vector Machine (SVM)* dan *K-NEAREST NEIGHBOR*
Hasil Penelitian	Dari hasil pengujian, hasil pengujian terbaik yaitu menggunakan metode (SVM) dan dengan fitur N-Gram Unigram didapatkan sebuah akurasi sebesar 87%%, spesifisitas 89%%, sensitivitas 76%, dan AUC 82%.

Tabel 2. 5 Penelitian Terdahulu 4

Penelitian 4	
Judul	ANALISIS SENTIMEN *REVIEW* BARANG BERBAHASA INDONESIA DENGAN METODE *SUPPORT VECTOR MACHINE* DAN *QUERY EXPANSION*
Peneliti & Tahun Penelitian	Dimas Joko Haryanto, Lailil Muflikhah, Mochammad Ali Fauzi (Universitas Brawijaya), Tahun 2018
Sumber	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer Vol.2, No.9
Tujuan	Untuk mengetahui review barang yang terdapat pada toko online dan mengelompokkannya menjadi sebuah umpan balik yang bernilai positif atau negative bagi toko tersebut.
Metode	Metode klasifikasi *Support Vector Machine (SVM)* tanpa *Query Expansion* dan *Support Vector Machine (SVM)* dengan *Query Expansion*
Hasil Penelitian	Dari hasil pengujian yang telah dilakukan, didapatkan sebuah perbandingan 2 akurasi dari metode SVM dengan *Query Expansion* dan dari SVM tanpa *Query Expansion.* Akurasi SVM dengan menggunakan *Query Expansion* yaitu 96.25%, akurasi ini lebih besar dibandingkan dengan akurasi SVM tanpa menggunakan *Query Expansion* yaitu 94.75%.

Tabel 2. 6 Penelitian Terdahulu 5

Penelitian 5	
Judul	IMPLEMENTASI ALGORITMA SUPPORT VECTOR MACHINE DAN CHI SQUARE UNTUK ANALISIS SENTIMEN USER FEEDBACK APLIKASI
Peneliti & Tahun Penelitian	Lulu Luthfiana, Julio Christian Young, dan Andre Rusli (Universitas Multimedia Nusantara), Tahun 2020
Sumber	ULTIMATICS, Vol. XII, No. 2
Tujuan	Untuk melakukan klasifikasi sentimen terhadap *user feedback* suatu aplikasi serta untuk mengetahui kinerja dari pengimplementasian algoritma SVM dan *Chi Square*.
Metode	*Support Vector Machine* dan *Chi Square*
Hasil Penelitian	Pengujian yang dilakukan pada penelitian ini didapatkan hasil yang terbaik yaitu dengan menggunakan metode algoritma *Support Vector Machine* dan dengan fitur seleksi *Chi Square*. Akurasi didapatkan sebesar 77%, *precision* 50%, *recall* 55%, dan *F1-Score* 73%

Tabel 2. 7 Penelitian Terdahulu 6

Penelitian 6	
Judul	*HIGHLIGHTING KEYPHRASES USING SENTI-SCORING AND FUZZY ENTROPY FOR UNSUPERVISED SENTIMENT ANALYSIS*
Peneliti & Tahun Penelitian	Srishti Vashishtha dan Seba Susan (Delhi Technological University), Tahun 2020
Sumber	*Expert System with Application*, Vol. 169
Tujuan	Untuk melakukan klasifikasi sentimen terhadap opini *online movie* serta melakukan eksperimen perbandingan hasil menggunakan Pang-Lee *Dataset* dan juga IMDB *Dataset*
Metode	*Senti-Scoring, Fuzzy Entropy, N-Gram Combination*
Hasil Penelitian	Sebelum menggunakan menggunakan *N-Gram Combination* hasil akurasi yang didapat dari Pang-Lee *Dataset* sebesar 49,6% sedangkan dari IMDB *Dataset* sebesar 48,75% Pengujian yang dilakukan pada penelitian ini didapatkan hasil yang terbaik yaitu dengan menggunakan menggunakan *N-Gram Combination* (*unigram-bigram-trigram*) dengan hasil akurasi sebesar 70%. Sedangkan dengan menggunakan *N-Gram Combination* pada IMDB *Dataset* didapatkan hasil akurasi sebesar 69,3%.

Dari beberapa penelitian terdahulu diatas memiliki keterkaitan dengan penelitian kali ini yang sedang dilakukan yaitu menganalisis sentimen dari opini atau *review* pengguna terhadap suatu produk atau jasa yang ditawarkan sebuah organisasi perusahaan. Selain itu juga dari penelitian terdahulu ada beberapa hal yang diadopsi atau diambil yaitu penggunaan metode *Support Vector Machine* (SVM) [24], [23], [30], [31], pengadopsian terhadap penggunaan tahapan data preprocessing yaitu tokenizing, transform case, stemming, stopword removal [11] dan penggunaan fitur N-Gram (*Unigram, Bigram, dan Trigram*) [23], tetapi pada penelitian kali ini yang sedang dilakukan data dan jumlah data yang digunakan

berbeda dengan penelitian terdahulu, data yang digunakan pada penelitian kali ini berupa data opini yang diambil dari *Google Play Store* dengan jumlah data yang lebih banyak daripada penelitian terdahulu. Selain itu juga dalam penelitian kali ini menggunakan fitur TFIDF untuk melakukan pembobotan kata terhadap opini yang sudah dikumpulkan yaitu *TFIDF Vectorizer* [19] dan juga melakukan proses pengujian menggunakan SVM dengan jumlah bobot perbandingan antara data uji dan data latih yang berbeda-beda [28] serta memodifikasi penggunaan fitur N-Gram yaitu dengan tambahan *N-Gram Combination unigram-bigram* (1,2), *unigram-trigram* (1,3), dan *bigram-trigram* (2,3) [32].