



## SENTIMEN DATA ULASAN MENGGUNAKAN METODE NAIVE BAYES STUDI KASUS THE WUJIL RESORT & CONVENTIONS PADA SITUS TRIPADVISOR

Faris Febri Rahanto<sup>✉</sup>, Iqbal Kharisudin

Jurusan Matematika, FMIPA, Universitas Negeri Semarang, Indonesia

Gedung D7 Lt. 1, Kampus Sekaran Gunungpati, Semarang 50229

### Info Artikel

Sejarah Artikel:  
Diterima Ferbruari 2021  
Disetujui Juni 2021  
Dipublikasikan Juni 2021

*Keywords:*  
*Naive Bayes Classifier*  
*Sentiment Analysis*  
*Text Mining*  
*Web Scraping*

### Abstrak

Situs tripadvisor menyediakan informasi mengenai ulasan-ulasan pengunjung The Wujil Resort & Conventions. Setiap pengunjung dapat memberikan ulasan baik berupa kritik, saran, ataupun penilaian terhadap hotel. Banyaknya ulasan yang masuk, maka diperlukan sebuah teknik khusus untuk menggali informasi dari ulasan tersebut. Penelitian ini bertujuan untuk menganalisis sentimen data ulasan menggunakan metode *Naive Bayes Classifier*. *Web scraping* digunakan untuk mendapatkan data pada halaman *website* tripadvisor dengan mengumpulkan data ulasan pengunjung The Wujil Resort & Conventions. Proses pengklasifikasian data ulasan dilakukan dengan *machine learning* menggunakan metode *Naive Bayes Classifier*. Selanjutnya hasil klasifikasi dianalisis dengan metode *text mining*, konsep utamanya adalah dengan melakukan eksplorasi seluas-luasnya dan ekstraksi dengan data yang sangat banyak dan terus bertambah. Sehingga ditemukan sebuah fakta dan informasi yang dianggap penting dan dapat berguna untuk berbagai bidang keperluan. Klasifikasi dengan metode *Naive Bayes* menunjukkan tingkat akurasi sebesar 76,6%. Secara umum, dengan metode *text mining* diperoleh informasi bahwa lebih banyak pengunjung yang memberikan penilaian positif daripada pengunjung yang memberikan penilaian negatif.

### Abstract

*The tripadvisor site provides information on visitor reviews of The Wujil Resort & Conventions. Each visitor can provide a review in the form of criticism, suggestions, or ratings of the hotel. The number of incoming reviews, it requires a special technique to extract information from these reviews. This study aims to analyze the sentiment of the review data using the Naive Bayes Classifier method. Web scraping is used to get data online on website pages, namely collecting visitor review data for The Wujil Resort & Conventions. The process of classifying the review data will be carried out using machine learning using the Naive Bayes Classifier method. Furthermore, the classification results will be analyzed using the text mining method, the main concept is to carry out extensive exploration and extraction with a large and growing number of data. So that we find facts and information that are considered important and can be useful for various fields of purpose. The classification using the Naive Bayes method shows an accuracy rate of 76.6%. In general, with the text mining method, information is obtained that there are more visitors who give positive ratings than visitors who give negative ratings.*

### How to cite:

Rahanto, Faris Febri & Kharisudin, Iqbal. 2021. Analisis Sentimen Data Ulasan Menggunakan Metode Naive Bayes. *UNNES Journal of Mathematics*. 10(1):55-62.

## PENDAHULUAN

Meningkatnya kebutuhan akan informasi mendorong manusia untuk mengembangkan teknologi baru agar pengolahan data dan informasi dapat dilakukan dengan mudah dan cepat. Salah satu perkembangan teknologi adalah akan mempermudah dan mempercepat proses pengolahan data, mencari informasi dan sebagainya (Josi, Abdillah, & Suryayusra, 2014). Berbagai kemudahan dan manfaat yang diperoleh pelaku bisnis dari internet, di antaranya adalah bagi perusahaan jasa yang bergerak di bidang pariwisata seperti hotel, travel, pemesanan tiket, dan sebagainya (Suhartono, 2013). Eskalasi *web 2.0* dan teknologinya telah sangat mendukung interaksi dan kolaborasi pengguna di media sosial seperti blog, wiki, aplikasi *web*, dan situs komersial. Perkembangan berbagai bentuk ekspresi opini *online* oleh pengguna mengidentifikasi peluang baru dan sumber daya penting jika dimanfaatkan dengan tepat. Ini secara efektif diekspresikan dalam bidang ilmu data dan pembelajaran mesin (Elnagar, Lulu, & Einea, 2018).

Ulasan pelanggan *online* semakin banyak tersedia di berbagai produk dan layanan dan menganalisis sentimen ulasan pelanggan telah menjadi sangat bermanfaat dalam bisnis, di mana bisnis dapat melacak ulasan positif dan negatif yang membantu mereka mengukur kinerja mereka secara keseluruhan dan dapat memainkan peran kunci dalam mengukur penjualan dan meningkatkan strategi pemasaran bisnis juga (Al-Saqqa, Al-Naymat, & Awajan, 2018). Meskipun ulasan *online* memiliki struktur yang buruk, ulasan tersebut berguna untuk membedakan preferensi dan permintaan tamu serta dalam memahami apa yang membuat tamu kembali atau menghindari hotel (Sánchez-Franco, Navarro-García, & Rondán-Cataluña, 2019). Keragaman pendapat yang ada dalam data tekstual, yang disediakan oleh pengguna, menimbulkan kompleksitas yang tidak diinginkan karena pemrosesan data sebesar itu adalah tugas yang hampir mustahil bagi manusia. Untuk tujuan ini, ilmuwan komputer menyediakan beberapa alat / algoritme penambangan data yang dapat membantu pengguna mengekstrak informasi yang relevan dari sejumlah besar data (Ray, Garain, & Sarkar, 2021).

Dalam penelitian ini, penulis memfokuskan pada analisis sentimen dari ulasan pengunjung The Wujil Resort & Conventions yang ada di situs tripadvisor. Tripadvisor menyediakan berbagai fasilitas

yang dapat memudahkan pengunjung untuk memperoleh informasi secara rinci tentang hotel yang akan dikunjungi, baik lokasi, banyak kamar, dan berbagai fasilitas lain yang tersedia. Selain itu, situs tripadvisor menyediakan informasi mengenai ulasan-ulasan para wisatawan. Setiap pengunjung dapat memberikan ulasan baik berupa kritik, saran, ataupun penilaian terhadap hotel yang pernah dikunjungi.

Penelitian ini bukan merupakan penelitian yang pertama, namun sudah ada beberapa penelitian terkait dengan metode *Naive Bayes*. Penelitian yang dilakukan oleh Xhemali, J. Hinde, & G. Stone (2009) yang berkonsentrasi pada perbandingan tiga metode yaitu *Naive Bayes*, *Decision Tree*, dan *Neural Networks* untuk mengklasifikasikan halaman web kursus pelatihan. Hasil penelitian menunjukkan bahwa *Naive Bayes* memberikan kinerja terbaik untuk domain studinya. Penelitian ini penting dilakukan untuk mengetahui dan membandingkan metode *Naive Bayes Classifier* dengan metode klasifikasi lain. Tujuan dari penelitian ini adalah untuk mengetahui akurasi metode *Naive Bayes* dalam mengklasifikasikan ulasan pengunjung The Wujil Resort & Conventions.

## METODE

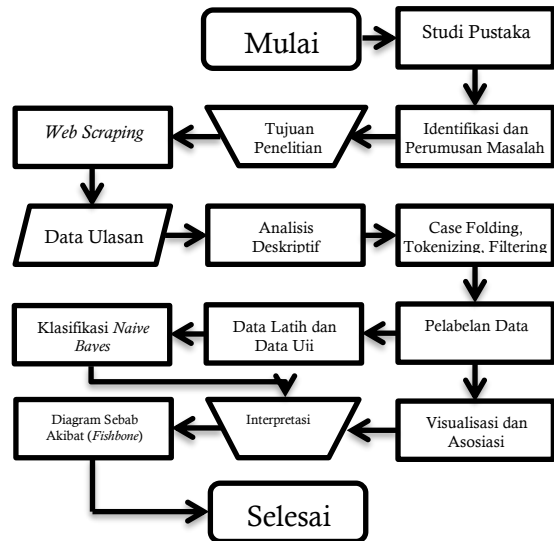
Metode Penelitian ini dilaksanakan dengan studi literatur, yaitu membaca dan menganalisis buku-buku dan jurnal-jurnal yang berkaitan dengan analisis sentimen, *Naive Bayes classifier*, dan *web scraping*. Tujuan dari metode studi literatur ini adalah untuk mendapatkan referensi, sehingga penelitian menjadi lebih mudah. Selanjutnya peneliti akan mengumpulkan data ulasan pengunjung The Wujil Resort & Conventions secara daring yang bersumber dari *website* tripadvisor. Data yang diambil berupa ulasan sejak bulan Agustus 2016 hingga bulan Februari 2020 sebanyak 188 ulasan.

Banyaknya ulasan yang masuk, maka diperlukan sebuah teknik khusus untuk mengumpulkan data secara daring dan sebuah analisis yang dapat membantu dalam memperoleh informasi dari data dalam skala besar. *Web scraping* menjadi solusi alternatif yang dapat digunakan untuk mendapatkan data secara daring pada halaman *website* sebelum data tersebut diolah dan dianalisis. *Web scraping* adalah proses pengambilan sebuah dokumen semi-terstruktur dari internet, umumnya berupa halaman web dalam bahasa *markup* seperti *HTML* atau *XHTML*, dan

menganalisis dokumen tersebut untuk diambil data (Turland, 2010). *Natural Language Processing (NLP)* digunakan untuk menyusun konsep analisis sentimen yang juga disebut sebagai penambangan opini atau *AI* emosi sebagai gagasan yang sangat luas yang akhirnya diidentifikasi sebagai aspek dari pelanggan. Kumpulan data yang digunakan dimanipulasi berdasarkan prosedur analisis tekstual serta linguistik komputasi untuk menggali informasi intrinsik yang tersedia (Kanna & Pandiaraja, 2019).

Proses pengklasifikasian data ulasan akan dilakukan dengan *machine learning* menggunakan metode *Naive Bayes Classifier*. *Naive Bayes Classifier* merupakan salah satu metode penalaran ketidakpastian yang menggunakan model probabilistik dan aturan Bayes untuk inferensi. *Naive Bayes* memaksakan asumsi naif bahwa di antara atribut-atribut yang diasumsikan independen secara bersyarat berdasarkan kelasnya. *Naive Bayes* juga dikenal memiliki performa tingkat tinggi dengan perhitungan yang sederhana (Mubarak, Adiwijaya, & Aldhi, 2017).

Selanjutnya hasil klasifikasi akan dianalisis dengan metode *text mining*. *Text mining* adalah tipe *natural-language processing* atau pengolahan bahasa alami yang menguraikan istilah (berupa kata dan frasa) dari dokumen tertentu (Gegick, Rotella, & Xie, 2010). Salah satu manfaat dari *text mining* adalah menghasilkan inovasi yang membantu orang untuk mengerti akan suatu sistem dengan menggunakan gudang dokumen (Srivastava & Sahami, 2009). *Text mining* atau *text analytics* adalah istilah yang mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur. Pada dasarnya, *text mining* merupakan bidang interdisiplin yang mengacu pada perolehan informasi (*information retrieval*), *data mining*, pembelajaran mesin (*machine learning*), statistik, dan komputasi linguistik (Han, Kamber, & Pei, 2012). Konsep utamanya adalah dengan melakukan eksplorasi seluas-luasnya dan ekstraksi dengan data yang sangat banyak dan terus bertambah. Sehingga ditemukan sebuah fakta dan informasi yang dianggap penting dan dapat berguna untuk berbagai bidang keperluan. Langkah akhir dari metode penelitian ini adalah penarikan kesimpulan. Tahapan yang dilakukan pada penelitian ini disajikan pada Gambar 1.



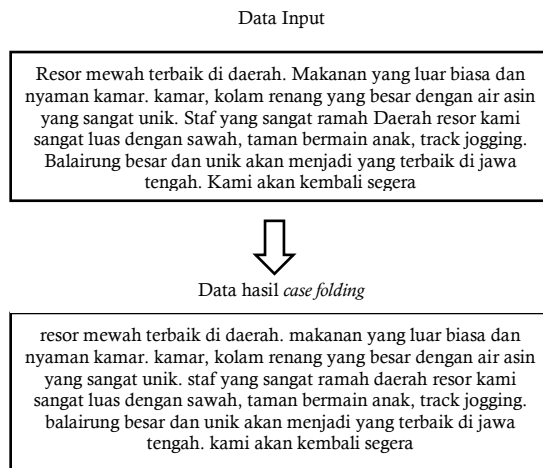
Gambar 1 Tahap penelitian

## HASIL DAN PEMBAHASAN

### *Text Preprocessing*

#### *Case Folding*

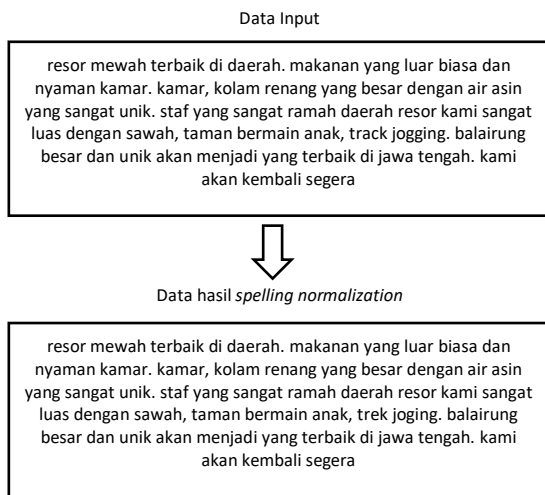
*Case folding* merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil, hanya huruf 'a' sampai dengan 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter (pembatas). Contoh tahap *case folding* dapat dilihat pada Gambar 2.



Gambar 2 Tahap *case folding*

#### *Spelling Normalization*

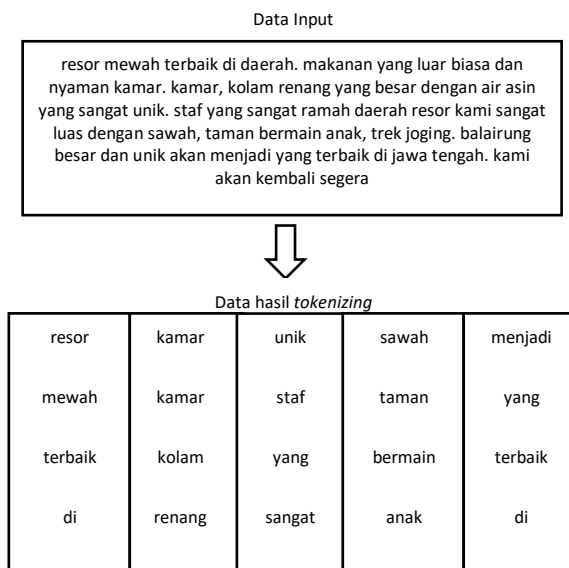
*Spelling normalization* merupakan perbaikan kata-kata yang salah eja atau disingkat dengan bentuk tertentu. Misalnya kata "dengan" memiliki banyak bentuk penulisan seperti dg, dgn, dan banyak kata lainnya. Contoh tahap *spelling normalization* dapat dilihat pada Gambar 3.



Gambar 3 Tahap *spelling normalization*

**Tokenizing**

*Tokenizing* merupakan proses pemisahan teks menjadi potongan kata yang disebut token. *Tokenizing* dilakukan untuk mendapatkan token atau potongan kata yang akan menjadi entitas yang memiliki nilai dalam penyusunan matriks dokumen pada tahap berikutnya. Contoh tahap *tokenizing* dapat dilihat pada Gambar 4.

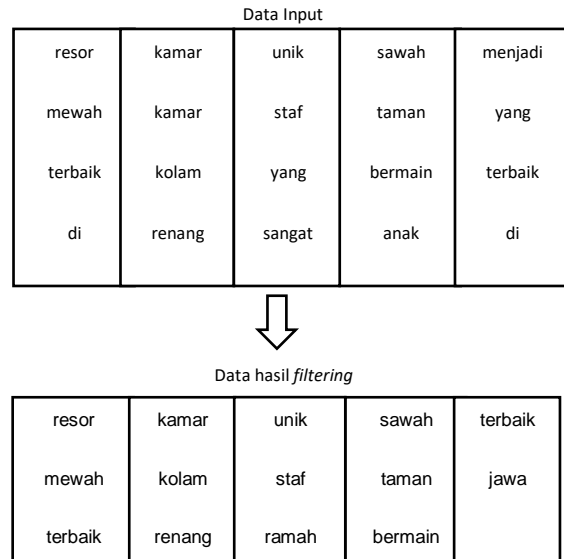


Gambar 4 Tahap *tokenizing*

**Filtering**

*Filtering* adalah proses mengambil kata-kata penting dari hasil *tokenizing*. Tahap *filtering* dapat menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stopword / stoplist* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopword* adalah “akan”, “dengan”, “di”, “yang”, dan lain-lain. Selain

dengan menggunakan *stopword*, tahap *filtering* juga dilakukan secara manual, yakni dengan menghapus kata-kata yang tidak terdapat dalam daftar *stopword* yang dianggap kurang penting dan kurang berpengaruh secara signifikan terhadap hasil analisis. Contoh penggunaan *filtering* dapat dilihat pada Gambar 5.



Gambar 5 Tahap *filtering*

**Pelabelan Kelas Sentimen**

Kalimat yang memiliki skor  $\geq 0$  akan diklasifikasikan ke dalam kelas positif, sedangkan kalimat yang memiliki skor  $< 0$  diklasifikasikan ke dalam kelas negatif. Perbandingan jumlah data hasil pelabelan kelas sentimen disajikan pada Tabel 1.

Tabel 1 Perbandingan banyak data pada kelas sentimen

| Sentimen | Banyak Ulasan |
|----------|---------------|
| Positif  | 152           |
| Negatif  | 36            |

Adapun rumus perhitungan skor sentimen yang digunakan dalam proses pelabelan adalah sebagai berikut:

$$\text{Skor} = \text{banyak kata positif} - \text{banyak kata negatif}$$

Simulasi perhitungan skor sentimen ulasan pengunjung The Wujil Resort & Conventions disajikan pada Tabel 2.

Tabel 2 Simulasi perhitungan skor sentimen

| Ulasan              | Kata Positif | Kata Negatif |
|---------------------|--------------|--------------|
| resor mewah terbaik | mewah        |              |
| daerah makanan      | terbaik      | asin         |
| nyaman kamar        | nyaman       |              |

|  |                                |   |
|--|--------------------------------|---|
| kamar kolam renang<br>besar air asin unik<br>staf ramah daerah<br>resor luas sawah<br>taman bermain trek<br>joging balairung<br>besar unik terbaik<br>jawa | besar<br>unik<br>ramah<br>luas |   |
| Jumlah   | 7                              | 1 |

Sehingga dengan demikian diperoleh perhitungan sebagai berikut:

$$Skor = \text{banyak kata positif} - \text{banyak kata negatif}$$

$$Skor = 7 - 1$$

$$Skor = 6$$

Skor akhir yang diperoleh dari simulasi perhitungan bernilai  $\geq 0$ , sehingga hasil klasifikasi ulasan adalah positif.

**Pembuatan Data Latih dan Data Uji**

Data latih digunakan oleh algoritma klasifikasi untuk membentuk sebuah model *classifier*, model ini merupakan representasi pengetahuan yang akan digunakan untuk prediksi kelas data baru yang belum pernah ada, semakin besar data latih yang digunakan, maka akan semakin baik *machine* dalam memahami pola data. Data uji digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar.

Tabel 3 Perbandingan data latih dan data uji

| Klsifikas<br>i | Jumlah | Data<br>Latih<br>(75%) | Data Uji<br>(25%) |
|----------------|--------|------------------------|-------------------|
| Positif        | 152    | 115                    | 37                |
| Negatif        | 36     | 26                     | 10                |
| Total          | 188    | 141                    | 47                |

**Klasifikasi Naive Bayes**

**Hasil Confusion Matrix Ulasan**

*Confusion matrix* digunakan untuk memudahkan dalam proses perhitungan akurasi dengan mengetahui jumlah data uji yang terklasifikasi dengan benar dan jumlah data uji yang salah pengklasifikasiannya. Hasil *confusion matrix* dari data ulasan dapat dilihat pada table berikut.

Tabel 4 Nilai *confusion matrix* ulasan data uji

| Predictio<br>n | Actual  |         |
|----------------|---------|---------|
|                | Negatif | Positif |

|         |    |    |
|---------|----|----|
| Negatif | 0  | 1  |
| Positif | 10 | 36 |

Pada Tabel 4, didapatkan hasil *confusion matrix* dari ulasan data uji. Jumlah *true positive* adalah 0, jumlah *true negative* adalah 36, jumlah *false positive* adalah 10, dan jumlah *false negative* adalah 1.

Untuk menghitung nilai *recall* atau *sensitivity* yaitu dengan menggunakan rumus:

$$Sensitivity (Recall) = \frac{TP}{TP + FN} = \frac{0}{0 + 10} = \frac{0}{10} = 0$$

Selanjutnya menghitung nilai *precision* yaitu dengan menggunakan rumus:

$$Precision = \frac{TP}{TP + FP} = \frac{0}{0 + 1} = \frac{0}{1} = 0$$

Selanjutnya menghitung nilai *specificity* yaitu dengan menggunakan rumus:

$$Specificity = \frac{TN}{TN + FP} = \frac{36}{36 + 1} = \frac{36}{37} = 0,97297$$

**Hasil Akurasi Ulasan**

Untuk menghitung nilai akurasi yaitu dengan menggunakan rumus:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{0 + 36}{0 + 36 + 1 + 10} = \frac{36}{47} = 0,7659574$$

Hasil yang didapatkan cukup bagus, yaitu bernilai 0,7659574.

**Visualisasi dan Asosiasi**







|           |      |             |      |
|-----------|------|-------------|------|
| Tidur     | 0,32 | Membayar    | 0,42 |
| Ditemukan | 0,22 | Restorannya | 0,42 |

Berdasarkan Tabel 6 diperoleh asosiasi antar kata pada ulasan negatif, kata-kata tersebut merupakan topik yang paling sering dibicarakan pengunjung dalam ulasannya. Kata-kata yang berasosiasi dengan kata “kamar” memberikan informasi bahwa pengunjung memberikan catatan kepada pengelola hotel agar staf hotel lebih teliti saat membersihkan kamar tidur, karena pengunjung menemukan bercak kotor sudut ruangan kamar tidur.

**Diagram Fishbone**



Gambar 10 Diagram *fishbone* komplain pengunjung

Berdasarkan Gambar 10 dapat diketahui bahwa faktor-faktor yang menyebabkan The Wujil Resort & Conventions memiliki ulasan negatif yaitu fasilitas, lingkungan, dan pelayanan.

**PENUTUP**

Teknik *web scraping* dapat digunakan sebagai cara alternatif untuk mengambil data dari halaman *website*. *Web scraping* mempermudah dan mempercepat proses pengambilan data dalam skala besar secara otomatis di internet. Hasil klasifikasi sentimen menggunakan metode *Naive Bayes* menghasilkan tingkat akurasi sebesar 76,6%, artinya dari 47 data ulasan yang diuji, terdapat 36 ulasan yang benar pengklasifikasiannya oleh metode *Naive Bayes*. Pengunjung The Wujil Resort & Conventions mayoritas memiliki persepsi yang baik terhadap hotel. Pengunjung banyak memberikan penilaian positif antarlain tentang penilaian kamar hotel bergaya modern minimalis dengan balkon, staf yang ramah dan membantu, menu restoran yang banyak variasi makanan, serta adanya fasilitas penunjang lain seperti taman bermain,

kolam renang, *wifi*, dan *jogging track*. Namun masih ada beberapa penilaian negatif pengunjung seperti pelayan pembersih kamar yang kurang teliti saat membersihkan sudut ruang kamar. Berdasarkan diagram *fishbone*, faktor yang harus ditingkatkan dalam pemecahan masalah hasil dari ulasan negatif adalah faktor fasilitas, pelayanan, dan lingkungan.

**DAFTAR PUSTAKA**

Al-Saqqqa, S., Al-Naymat, G., & Awajan, A. (2018). A large-scale sentiment data classification for online reviews under apache spark. *Procedia Computer Science*, 141, 183–189. <https://doi.org/10.1016/j.procs.2018.10.166>

Elnagar, A., Lulu, L., & Einea, O. (2018). An Annotated Huge Dataset for Standard and Colloquial Arabic Reviews for Subjective Sentiment Analysis. *Procedia Computer Science*, 142, 182–189. <https://doi.org/10.1016/j.procs.2018.10.474>

Gegick, M., Rotella, P., & Xie, T. (2010). Identifying security bug reports via text mining: An industrial case study. *Proceedings - International Conference on Software Engineering*, (June 2010), 11–20.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Third Edit). Waltham, MA: Morgan Kaufmann.

Josi, A., Abdillah, L. A., & Suryayusra. (2014). Penerapan Teknik *Web Scraping* pada Mesin Pencari Artikel Ilmiah. 159–164.

Kanna, P. R., & Pandiaraja, P. (2019). An Efficient Sentiment Analysis Approach for Product Review using Turney Algorithm. *Procedia Computer Science*, 165(2019), 356–362. <https://doi.org/10.1016/j.procs.2020.01.038>

Mubarok, M. S., Adiwijaya, A., & Aldhi, M. D. (2017). Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes. *AIP Conference Proceedings*, 1867, 020060. <https://doi.org/10.1063/1.4994463>

- Ray, B., Garain, A., & Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, 98, 106935. <https://doi.org/10.1016/j.asoc.2020.106935>
- Sánchez-Franco, M. J., Navarro-García, A., & Rondán-Cataluña, F. J. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research*, 101(December), 499–506. <https://doi.org/10.1016/j.jbusres.2018.12.051>
- Srivastava, A. N., & Sahami, M. (2009). *Text Mining Classification, Clustering, and Applications*. Boca Raton: CRC Press.
- Suhartono, B. (2013). Fungsi dan Manfaat Internet dalam Bidang Bisnis dan Perdagangan. Retrieved September 15, 2019, from <https://bambangsuhartono.wordpress.com/2013/02/04/fungsi-dan-manfaat-internet-dalam-bidang-bisnis-dan-perdagangan/>
- Turland, M. (2010). *Guide to Web Scraping with PHP* (First Edit). Toronto: Marco Tabini & Associates, Inc.
- Xhemali, D., J. Hinde, C., & G. Stone, R. (2009). Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science*, 4(1), 16–23.