

Analisis Sentimen Ulasan *Customer* Kopi TMLST Menggunakan Algoritma *Naïve Bayes*

Dhiya Azizah Hamidah¹, Ratna Salkiawati¹, Rafika Sari^{1,*}

* Korespondensi: rafika.sari@dsn.ubharajaya.ac.id

¹ Informatika; Fakultas Ilmu Komputer, Universitas Bhayangkara Jakarta Raya; Jl. Raya Perjuangan No.81, Margamulya, Bekasi Utara, Kota Bekasi, Jawa Barat; dhiya.azizah.hamidah19@mhs.ubharajaya.ac.id, ratna.tind@dsn.ubharajaya.ac.id, rafika.sari@dsn.ubharajaya.ac.id

Submitted : 12 Maret 2024
Revised : 26 Maret 2024
Accepted : 23 April 2024
Published : 31 Mei 2024

Abstract

The rapid development of Coffeeshop is currently influenced by advances in internet technology, the existence of online food applications and websites, such as Shopeefood, and Google Maps, can help people place online orders that have no time limit. However, there are problems that arise over time such as, in collecting feedback from customers the more review data available on Google Maps and online food applications, namely Shopeefood. Therefore, a solution is needed that can help TMLST Coffee to collect, process, and analyze feedback from customers on online food applications such as Shopeefood and Google Maps in a better and more structured manner. In this study, retrieving and collecting customer review data was carried out using web scrapping techniques taken through online food applications, namely Shopeefood and Google Maps, but collecting review data was also carried out by distributing questionnaires via google forms filled out by TMLST Coffee customers. Furthermore, the method used in this research is Naive Bayes which aims as a classification method and is able to classify customer comments into positif or negatif. And review data processing is done using the Cross-Industry Standard Process for Data Mining (CRIPS-DM) method. The CRIPS-DM stage involves the research and implementation process of the stages that have been carried out previously. The results of this study produce a high level of accuracy in predicting positif and negatif sentiment, with an accuracy of 0.82 or 82%. In addition, it produces a positif recall of 0.76 or 76% and a negatif recall of 0.89 or 89%. indicating that the model has a good ability to identify correctly. With the evaluation results of the model used, it gives an indication that Naive Bayes can be an effective choice in conducting sentiment analysis on TMLST Coffee review data.

Keywords: CRIPS-DM, Feedback, Naive Bayes Algorithm, Sentiment Analysis

Abstrak

Perkembangan *Coffeeshop* yang pesat saat ini ikut berpengaruh dengan adanya kemajuan teknologi internet, adanya aplikasi *online food* dan situs web, seperti *Shopeefood*, dan *Google Maps*, dapat membantu masyarakat dalam melakukan pemesanan *online* yang tidak memiliki batas waktu. Namun, adanya permasalahan yang muncul seiring berjalannya waktu seperti, dalam mengumpulkan *feedback* dari *customer* semakin banyaknya data ulasan yang tersedia di *Google Maps* dan aplikasi *online food* yaitu *Shopeefood*. Oleh karena itu, diperlukan suatu solusi yang dapat membantu TMLST Coffee untuk mengumpulkan, memproses, dan menganalisis *feedback* dari *customer* pada aplikasi *online food* seperti *Shopeefood* dan *Google Maps* dengan lebih baik dan terstruktur. Dalam penelitian ini pengambilan dan mengumpulkan data ulasan *customer* dilakukan dengan teknik *web scrapping* yang diambil melalui aplikasi *online food* yaitu *Shopeefood* dan *Google Maps*,

akan tetapi pengumpulan data ulasan dilakukan juga dengan menyebar kuesioner melalui *google form* yang diisi oleh *customer TMLST Coffee*. Selanjutnya, metode yang digunakan dalam penelitian ini adalah *Naïve Bayes* bertujuan sebagai metode klasifikasi dan mampu mengelompokkan komentar customer menjadi positif ataupun negatif. Dan pengolahan data ulasan dilakukan dengan menggunakan metode *Cross-Industry Standard Process for Data Mining (CRIPS-DM)*. Tahap *CRIPS-DM* melibatkan proses penelitian dan pelaksanaan dari tahap-tahap yang telah dilakukan sebelumnya. Hasil dari penelitian tingkat akurasi yang tinggi dalam memprediksi sentimen positif dan negatif, dengan akurasi sebesar 0.82 atau 82%. Selain itu, menghasilkan *recall* positif sebesar 0.76 atau 76% dan *recall* negatif sebesar 0.89 atau 89%. menunjukkan bahwa model memiliki kemampuan yang baik dalam mengidentifikasi dengan benar. Dengan hasil evaluasi model yang digunakan memberikan indikasi bahwa *Naïve Bayes* dapat menjadi pilihan yang efektif dalam melakukan analisis sentimen pada data ulasan *TMLST Coffee*.

Kata kunci: Algoritma *Naïve Bayes*, Analisis Sentimen, *CRIPS-DM*, Ulasan *Customer*.

1. Pendahuluan

Perkembangan *Coffeeshop* yang pesat saat ini ikut berpengaruh dengan adanya kemajuan teknologi internet (Agustina et al., 2022), adanya aplikasi *online food* dan situs web, seperti *Shopeefood*, dan *Google Maps*, dapat membantu masyarakat dalam melakukan pemesanan *online* yang tidak memiliki batas waktu serta dapat membantu para pelaku usaha dalam mencapai target pasaran. Sementara itu, di dalam layanan aplikasi *online food* dan situs web seperti *Shopeefood* dan *Google Maps* terdapat fitur untuk mencari informasi tentang bisnis *Coffeeshop*, selanjutnya terdapat juga fitur ulasan dan foto *coffeshop* tertentu.

Salah satu *coffeshop* yang ikut berpengaruh dengan adanya kemajuan teknologi internet adalah *TMLST Coffee*. Permasalahan yang muncul seiring berjalannya waktu seperti, dalam mengumpulkan *feedback* dari *customer* semakin banyaknya data ulasan yang tersedia di *Google Maps* dan aplikasi *online food* yaitu *Shopeefood*. Akibatnya, membuat *TMLST Coffee* kesulitan untuk memantau dan mengevaluasi setiap ulasan yang diberikan oleh *customer* secara efektif dan efisien. Ulasan dari *customer* tersebut biasanya menjadi tolak ukur yang efektif dan efisien untuk menemukan informasi dan akan berpengaruh kepada calon *customer* baru terhadap *Coffeeshop* tersebut. Tujuan dari penelitian adalah untuk mengetahui *feedback* dari *customer* dan menganalisis ulasan *customer TMLST Coffee* yang telah membeli kopi dari aplikasi *online food* serta pada *Google Maps*.

Untuk mengumpulkan dan menganalisis ulasan *customer* yang begitu banyak, akan sangat membutuhkan waktu yang lama untuk membaca satu persatu ulasan tersebut, maka diperlukan teknik khusus dengan menggunakan teknik *scrapping*. Setelah dilakukan pengambilan dan mengumpulkan data ulasan, maka dilakukan analisis sentimen dengan menggunakan metode algoritma *Naïve Bayes* (Mutawalli et al., 2022; Yusuf & Sari, 2022).

Metode *Naïve Bayes* merupakan salah satu metode yang dapat digunakan pada analisis sentimen karena algoritma ini bertujuan sebagai metode klasifikasi dan mampu mengelompokkan komentar *customer* menjadi positif ataupun negatif. Dan metode *Naïve Bayes* memiliki tingkat akurasi yang cukup unggul dibandingkan dengan algoritma yang lainnya (Estika et al., 2021). Dalam penelitian ini, data yang diambil merupakan data ulasan *customer* pada aplikasi *online food* yaitu *Shopeefood* dan *Google Maps* serta dilakukan juga dengan menyebarkan kuesioner kepada *customer Coffeeshop TMLST Coffee*.

2. Metode Penelitian

2.1. Data Penelitian

Dalam melakukan penelitian pengambilan dan mengumpulkan data primer yang merupakan ulasan *customer* diperoleh melalui dua sumber yaitu: data kuesioner pelanggan dan data ulasan dari aplikasi *online food* yaitu *Shopeefood* dan *Google Maps*. Kemudian pada tahap *Scrapping Data* menggunakan teknik *web scrapping*. Data yang digunakan diambil dari data ulasan *customer* aplikasi *online food* yaitu *Shopeefood* dan *Google Maps* pada rentang waktu yang telah ditentukan.

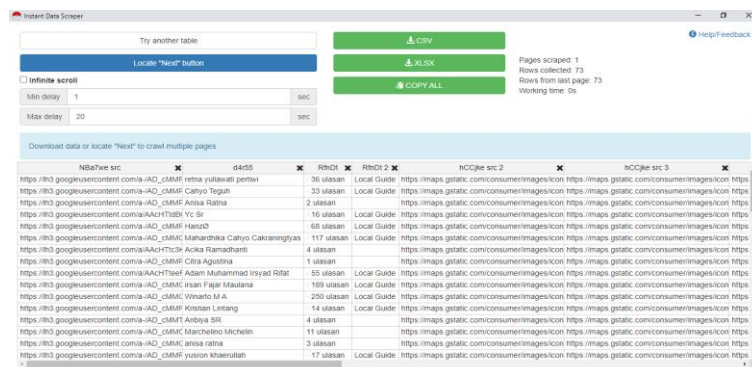
2.2. Metode Data Mining

Metode *data mining* yang dilakukan pada penelitian yaitu menggunakan metode *Cross-Industry Standard Process for Data Mining (CRIPS-DM)* (Kurniawan & Yasir, 2022; Putro & Sari, 2023). Tahap ini melibatkan proses penelitian dan pelaksanaan dari tahap-tahap yang telah dilakukan sebelumnya, yang telah dijelaskan sebelumnya, dengan tahapannya yaitu: (a) *Business Understanding*; (b) *Data understanding*; (c) *Data preparation*; (d) *Modeling*; (e) *Evaluation*; dan (f) *Deployment*.

3. Hasil dan Pembahasan

3.1. Data Preparation

Hasil pengambilan data ulasan *customer TMLST Coffee* pada aplikasi *online food* yaitu *Shopeefood* dan pada *Google Maps* dengan menggunakan teknik *Scrapping* dengan *Instant Data Scraper* versi 1.0.8 dan menyimpan data dalam format *Comma Separate Value (CSV)* yang diperlihatkan oleh Gambar 1.



Sumber: Hasil Penelitian (2023)

Gambar 1. *Scrapping* Ulasan Customer dengan *Instant Data Scraper*

a. Memuat *Dataset*

Pada Gambar 2 memuat *Dataset* menjelaskan bahwa `df = pd.read_csv('Data/Scrap_TMLST_Coffee.csv')`: untuk membaca file csv yang bernama "Scrap_TMLST_Coffee.csv" pada folder "Data" dan menyimpan data pada file tersebut ke dalam variabel `df`.

```
[3] 1 df = pd.read_csv('Scrap_TMLST_Coffee.csv')
```

Sumber: Hasil Penelitian (2023)

Gambar 2. Memuat *Dataset*

b. Data Ulasan *Customer*

Setelah dilakukan pengambilan dan pengumpulan data ulasan, tahap berikutnya adalah tahap pengolahan data ulasan. Gambar 3 data ulasan lima baris teratas yang sudah melalui tahapan pengolahan data ulasan.

	Name	Komen
0	retna yulianiwati pertiwi	Tempat nongkrong baru buat gw sm suami dan anak2. Tempat yg nyaman lah, luas jg, kopinya best. camilan nya jg anak2 suka sama es cokelat dan greentea nya juarak. 🍷...
1	Yc Sr	Kopinya enak...suasana outdoor cocok banget buat nongki...minuman coklatnya aq suka banget...dessert nya jg...sayang aq lupa foto pas nongki...next deh fotonya. gelasnya di aj...
2	Acika Ramadhanti	kopinya enak bgt dan makanannya juga oke, tempatnya outdoor jadi adem banget, baristanya juga ramah2
3	Anisa Ratna	enak tempatnya, nyaman buat ngobrol dan nongkrong lama2
4	Mahardhika Cahyo Cakraningtyas	Kopinya enak, minusnya kalo malem banyak nyamuk aja

Sumber: Hasil Penelitian (2023)

Gambar 3. Data Ulasan Lima Baris Teratas

Selanjutnya untuk mengetahui informasi tentang *dataset*, dapat digunakan metode `df.info()`. Dari hasil Gambar 3, dapat diamati bahwa *dataset* ini terdiri dari berbagai kolom dan baris, serta tipe data. Dapat dilihat bahwa tipe data secara keseluruhan adalah objek, dengan kolom Name, Komentar, dan Sentimen sebagai komponennya. Dapat dilihat Gambar 4 Informasi *Dataset*.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 154 entries, 0 to 153
Data columns (total 2 columns):
#   Column  Non-Null Count  Dtype
---  -
0   Name    154 non-null     object
1   Komen   154 non-null     object
dtypes: object(2)
memory usage: 2.5+ KB
```

Sumber: Hasil Penelitian (2023)

Gambar 4. Informasi *Dataset*

Pada Gambar 5 Missing Value *Dataset* yang menjelaskan hasil dari kode tersebut adalah jumlah nilai *null* atau *missing value* pada setiap kolom dari *dataset Scrap_TMLST*. Dari hasil Gambar 5 dapat disimpulkan bahwa tidak ada *missing value* pada *dataset Scrap_TMLST*.

```
Name      0  
Komen     0  
dtype: int64
```

Sumber: Hasil Penelitian (2023)

Gambar 5. *Missing Value*

c. *Data Preparation*

Proses Persiapan Data atau *Data Preparation* melibatkan serangkaian langkah yang dilakukan oleh seorang peneliti guna mengubah data mentah menjadi data yang berkualitas (Han et al., 2012). Proses ini melibatkan beberapa langkah yang perlu dilakukan, salah satunya pengecekan jumlah kata. Hasil pengecekan jumlah kata dapat dilihat pada Gambar 6, nantinya jumlah kata akan digunakan untuk *rule base* yang bertujuan untuk mengetahui kemunculan setiap kata dalam komentar yang kemudian diurutkan berdasarkan frekuensi kata dari yang paling sering muncul ke yang paling jarang. *Library Python* yang digunakan yaitu *Pandas* dan modul *Counter* dari *Python*. *Rule base* akan memiliki dua variabel yang berisi daftar kata kunci positif (*positif_keywords*) dan kata kunci negatif (*negatif_keywords*).

```
Kata: enak  
Jumlah: 35  
-----  
Kata: kopinya  
Jumlah: 32  
-----  
Kata: tempatnya  
Jumlah: 22  
-----  
Kata: coffee  
Jumlah: 16  
-----  
Kata: nyamuk  
Jumlah: 13  
-----  
Kata: tmlst  
Jumlah: 13  
-----
```

Sumber: Hasil Penelitian (2023)

Gambar 6. Mengecek Jumlah Kata

3.2. **Analisa Data**

a. *Rule Base*

Dalam tahapan ini, *rule base* digunakan untuk melakukan analisis sentimen pada sebuah teks. Dalam *rule base* memiliki fungsi 'get_sentimen(text) untuk menerima parameter text yang merupakan teks yang akan dianalisis sentimennya. Didalam *rule base* terdapat dua variabel yang berisi daftar kata kunci positif (*positif_keywords*) dan kata kunci negatif (*negatif_keywords*). Untuk setiap kata dalam "*positif_keywords*". Jika kata tersebut ditemukan dalam teks, maka akan menjadi nilai "Positif" sebagai hasil analisis sentimen. Selanjutnya, jika tidak ada kata-kata positif yang ditemukan, maka akan dilanjutkan untuk melakukan pengecekan terhadap kata-kata negatif. Untuk setiap kata dalam "*negatif_keywords*". Jika kata tersebut ditemukan dalam teks, maka akan mengganti menjadi nilai "Negatif" sebagai hasil analisis sentimen.

Jika tidak ada kata-kata positif maupun negatif yang ditemukan, maka akan mengembalikan menjadi nilai “Negatif” sebagai hasil analisis sentimen. Hal ini menunjukkan bahwa jika tidak ada indikasi kuat mengenai sentimen positif, maka sentimen akan diklasifikasikan sebagai negatif.

b. Mengubah Kolom Sentimen

	Name	Komen	Sentimen
0	retna yullawati pertivi	Tempat nongkrong baru buat gw sm suami dan anak2. \nTempat yg nyaman lah. luas jg. kopinya best. camilan nya jg anak2 suka sama es cokelat dan greentea nya juarak... 🍌 ...	1
1	Yc Sr	Kopinya enak...suasana outdoor cocok banget buat nongki...minuman coklatnya aq suka banget...dessert nya jg...sayang aq lupa foto pas nongki...next deh fotonya...gelasnya di aj... 🍌 ...	1
2	Azika Ramadhanti	kopinya enak bgt dan makanannya juga oke, tempatnya outdoor jadi adem banget, baristanya juga ramah2	1
3	Anisa Ratna	enak tempatnya, nyaman buat ngobrol dan nongkrong lama2	1
4	Mahardhika Cahyo Cakraningtyas	Kopinya enak, minusnya kalo malem banyak nyamuk aja	1

	Komen
0	Tempat nongkrong baru buat gw sm suami dan anak2. \nTempat yg nyaman lah. luas jg. kopinya best. camilan nya jg anak2 suka sama es cokelat dan greentea nya juarak... 🍌 ...
1	Kopinya enak...suasana outdoor cocok banget buat nongki...minuman coklatnya aq suka banget...dessert nya jg...sayang aq lupa foto pas nongki...next deh fotonya...gelasnya di aj... 🍌 ...
2	kopinya enak bgt dan makanannya juga oke, tempatnya outdoor jadi adem banget, baristanya juga ramah2
3	enak tempatnya, nyaman buat ngobrol dan nongkrong lama2
4	Kopinya enak, minusnya kalo malem banyak nyamuk aja

Sumber: Hasil Penelitian (2023)

Gambar 7. Mengubah Kolom Sentimen

Dari Gambar 7 Mengubah kolom sentimen menjelaskan bahwa `df['Sentimen'] = df['Sentimen'].map({'Positif':1,'Negatif':0})`, pada baris ini, kolom 'Sentimen' dalam DataFrame df diubah nilai-nilainya menggunakan metode `map()`. Nilai 'Positif' akan diubah menjadi angka 1, sedangkan nilai 'Negatif' akan diubah menjadi angka 0. Kode tersebut bertujuan untuk mengubah nilai-nilai kategori dalam kolom 'Sentimen' menjadi representasi numerik, yaitu 'Positif' menjadi 1 dan 'Negatif' menjadi 0. `df.head()`: Baris ini digunakan untuk menampilkan beberapa baris pertama dari DataFrame df setelah kolom 'Sentimen' diubah. Tujuannya adalah untuk memverifikasi perubahan yang telah dilakukan pada kolom tersebut.

c. Membagi Data

Gambar 8 Membagi Data X dan Y menjelaskan bahwa `X = df['Komen']`: Pada baris ini, variabel X dibuat dan diisi dengan kolom 'Komen' dari DataFrame df. Kolom 'Komen' mengandung teks atau komentar yang akan digunakan sebagai fitur dalam analisis sentimen. Sedangkan `y = df['Sentimen']`: Baris ini membuat variabel y dan mengisinya dengan kolom 'Sentimen' dari DataFrame df. Kolom 'Sentimen' berisi label atau target klasifikasi yang menunjukkan sentimen positif atau negatif dari setiap komentar dalam dataset.

```
[ ] 1 # Bagi data X dan y
     2 X = df['Komen']
     3 y = df['Sentimen']
```

Sumber: Hasil Penelitian (2023)

Gambar 8. Membagi Data X dan Y

d. Membaca *File Stopword*

Dalam tahapan ini Gambar 9 membaca *file stopwords* digunakan untuk membaca file teks yang berisi daftar kata yang merupakan *stopword* (kata-kata umum yang umumnya diabaikan dalam analisis teks).

```
1 f=open('stopwords.txt')
2 ina_stopword = f.read()
3 ina_stopword = ina_stopword.split('\n')
4 print('Stopwords 5 Kata Pertama', ina_stopword[:5])
```

Sumber: Hasil Penelitian (2023)

Gambar 9. Membaca *File Stopword*

e. Melakukan Pra-pemrosesan

Dalam tahap pra-pemrosesan dengan *stemming* digunakan untuk melakukan pra-pemrosesan teks pada Variabel X dengan menggunakan *stemming*.

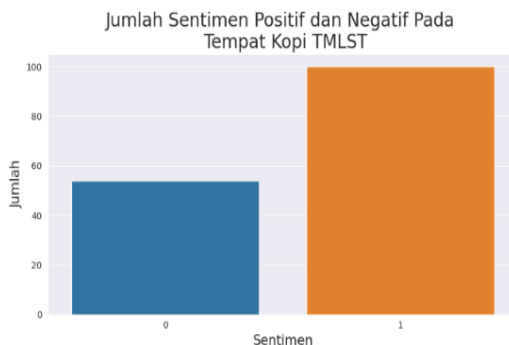
f. Hasil Pra-Pemrosesan

	Komen
0	nongkrong nyaman luas kopi best camil suka coklat greentea juarak
1	kopi enak suasana outdoor banget nongki minum coklat suka banget dessert nongki foto gelas
2	kopi enak outdoor banget baristanya ramah
3	enak nyaman ngobrol nongkrong
4	kopi enak minus nyamuk

Sumber: Hasil Penelitian (2023)

Gambar 10. Hasil Sesudah Pra-Pemrosesan

Gambar 10 menunjukkan hasil sesudah pra-pemrosesan menggunakan *stemming* pada Variabel X. Setelah proses *stemming* selesai, komentar-komentar tersebut telah diubah menjadi tipe *data frame*, yaitu *komentar_preprocessed*. Tipe *data frame* berisi komentar-komentar yang telah diproses dan siap digunakan untuk analisis selanjutnya. Pra-pemrosesan bertujuan untuk mengubah kata-kata dalam komentar menjadi bentuk dasar atau akar kata, sehingga mempermudah dalam analisis sentimen. Melalui tahapan ini memvisualisasi data dengan bentuk grafik yang dapat dilihat di Gambar 11 terdapat sejumlah ulasan yang telah diberikan oleh customer mengenai TMLST *Coffee*.



Sumber: Hasil Penelitian (2023)

Gambar 11. Jumlah Sentimen Positif dan Negatif

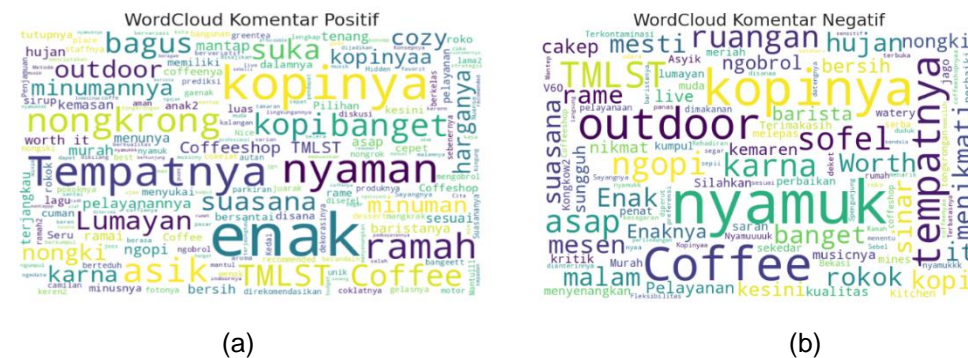
Dalam analisis sentimen, ulasan-ulasan tersebut telah dikategorikan menjadi dua klasifikasi utama, yaitu ulasan positif dan ulasan negatif, berdasarkan konten yang terkandung di dalamnya. Dari hasil analisis, terlihat bahwa terdapat sekitar 100 ulasan positif yang diberikan oleh customer, sedangkan jumlah ulasan negatif mencapai 54 ulasan. Dengan data ini, dapat disimpulkan bahwa mayoritas customer mengalami pengalaman yang positif ketika mengunjungi TMLST Coffee. Pada Gambar 11 dijelaskan perbandingan jumlah antara ulasan yang mengandung sentimen positif dan negatif. Sentimen positif diinisialisasikan dengan tanda angka 1 sedangkan untuk sentimen negatif diinisialisasikan dengan tanda angka 0.



Sumber: Hasil Penelitian (2023)

Gambar 12. Wordcloud

Selain memvisualisasikan dengan grafik, dilakukan juga dengan menggunakan wordcloud yang dapat dilihat di Gambar 12, *wordcloud* ialah tampilan kata-kata yang paling sering muncul akan ditampilkan dengan ukuran yang lebih besar, sedangkan kata-kata yang jarang muncul akan ditampilkan dengan ukuran yang lebih kecil (Widodo, 2021). Hal ini membantu untuk dengan cepat melihat kata-kata kunci yang sering terkait dengan TMLST Coffee dan mencerminkan sentimen *customer*.



Sumber: Hasil Penelitian (2023)

Gambar 13. (a) Wordcloud Positif dan (b) Wordcloud Negatif

Hasil wordcloud positif seperti Gambar 13 memvisualisasikan kata yang sering muncul pada label positif memperlihatkan bahwa, sebagian besar ulasan berlabel positif memuat kata-kata seperti “nyaman”, “asik”, “ramah”, dan “enak”. Hal ini menunjukkan bahwa customer banyak memberikan ulasan positif terkait dengan cita rasa kopi, kualitas layanan, dan suasana kafe

yang menyenangkan. Hasil wordcloud negatif dari Gambar 13 memvisualisasikan kata-kata yang sering muncul pada label negatif memperlihatkan bahwa, sebagian besar ulasan berlabel negatif memuat seperti “nyamuk”, “hujan”, “sofel”, “roko” atau “asap”, mungkin juga muncul dengan ukuran yang lebih kecil, menunjukkan adanya beberapa aspek yang perlu diperbaiki atau kekhawatiran yang perlu diperhatikan.

g. TF-IDF

Sebelum melakukan analisis sentimen dengan menggunakan algoritma *Naïve Bayes*, dilakukan ekstraksi fitur dengan menggunakan TF-IDF (Estika et al., 2021). Data kemudian dibagi menjadi data training dan data testing dengan rasio 80:20, yang berarti 80% dari jumlah data akan digunakan sebagai data *training* dan 20% dari jumlah data akan digunakan sebagai data *testing*. Pada Gambar 13 kode menggunakan algoritma TF-IDF (*Term Frequency-Inverse Document Frequency*) untuk mengubah teks ulasan customer yang telah di preproses menjadi representasi numerik.

```
array(['asap', 'asap roko', 'asap rokok', 'asik', 'asik murah'],  
      dtype=object)
```

Sumber: Hasil Penelitian (2023)

Gambar 14. TIF-IDF

h. *Imbalance Data*

```
1 y.value_counts()  
  
1    100  
0     54  
Name: Sentimen, dtype: int64
```

Sumber: Hasil Penelitian (2023)

Gambar 15. *Imbalance Data*

Berdasarkan dari Gambar 15 jumlah ulasan positif dan negatif di atas, terlihat bahwa distribusi ulasan tidak seimbang antara ulasan negatif dan positif. Ulasan positif memiliki jumlah yang lebih banyak daripada ulasan negatif.

```
1 smote = SMOTE()  
2 X_tf_idf, y = smote.fit_resample(X_tf_idf, y)  
3 print(sorted(Counter(y).items()))  
  
[(0, 100), (1, 100)]
```

Sumber: Hasil Penelitian (2023)

Gambar 16. *Handling Imbalance Data*

Gambar 16 *Handling Imbalance Data* menjelaskan langkah-langkah untuk menangani ketidakseimbangan data menggunakan metode *SMOTE* (*Synthetic Minority Over-sampling Technique*).

i. Normalisasi

Normalisasi data bertujuan untuk mengubah skala data ke dalam rentang yang spesifik, dalam hal ini rentang antara 0 dan 1.

j. *Splitting Data*

Splitting data digunakan untuk membagi *dataset* menjadi data latih (*train*) dan data uji (*test*) menggunakan fungsi `train_test_split()`. Tujuan dari pemisahan adalah untuk menguji performa model pada data yang belum pernah dilihat sebelumnya.

```
Jumlah Data Training X (160, 115)
Jumlah Data Training Y (160,)
=====
Jumlah Data Test X (40, 115)
Jumlah Data Test Y (40,)
```

Sumber: Hasil Penelitian (2023)

Gambar 17. Jumlah Data Latih dan Data Uji

Dapat dilihat pada Gambar 17 Jumlah Data Latih dan Data uji, dimana jumlah data *training* (latih) berjumlah 160 baris dan 115 kolom, sedangkan untuk jumlah data testing (uji) berjumlah 40 baris dan 115 kolom. Untuk 115 kolom ini didapatkan dari proses TF-IDF. Selanjutnya, dengan melakukan pemisahan data menjadi data latih dan data uji, model dapat melatih dirinya menggunakan data latih untuk kemudian menguji kinerja model menggunakan data uji. Tujuan dari pengujian adalah untuk mengevaluasi sejauh mana model dapat melakukan prediksi dengan tingkat akurasi yang baik.

3.3. Modeling

Dalam tahap ini, terjadi pengembangan model yang bertujuan untuk melakukan prediksi sentimen positif dan negatif berdasarkan data yang telah melalui tahap sebelumnya. Model tersebut dibuat dan dilatih dengan menggunakan data training untuk mempelajari pola dan hubungan antara fitur-fitur data dengan label sentimen. Setelah model selesai dikembangkan, dilakukan evaluasi performanya dengan menggunakan data testing. Pada Gambar 18 adalah *flowchart* algoritma *Naïve Bayes*.

```
1 model = GaussianNB()
2 model.fit(X_train, y_train)
```

▾ GaussianNB
GaussianNB()

Sumber: Hasil Penelitian (2023)

Gambar 18. Model *Naïve Bayes*

Pada Gambar 18 Model *Naïve Bayes*, sebuah model *Gaussian Naïve Bayes* (*GaussianNB*) diinisialisasi dan dilatih menggunakan data latih (*X_train* dan *y_train*). Model ini akan belajar dari data latih untuk mempelajari hubungan antara fitur-fitur yang ada dan label yang sesuai. Setelah pelatihan selesai, model siap untuk melakukan prediksi berdasarkan komentar pada data uji.

```
1 pred = model.predict(X_test)
2 pred[:5]
```

array([0, 0, 0, 0, 0])

Sumber: Hasil Penelitian (2023)

Gambar 19. *Predict Model Naïve Bayes*

Pada Gambar 19 *Predict Model Naïve Bayes*, dilakukan prediksi sentimen menggunakan model yang telah dibuat sebelumnya. Data yang digunakan untuk prediksi adalah data *testing* (*X_test*). Setelah melakukan prediksi, hasil prediksi tersebut disimpan dalam variabel 'pred'. Kemudian, dengan memanggil 'pred[:5]', dapat melihat lima hasil prediksi pertama yang diperoleh dari model. Hasil prediksi ini merupakan label sentimen yang diprediksi oleh model untuk masing-masing sampel data dalam data *testing*.

3.4. Evaluasi

Didalam tahap evaluasi bertujuan untuk menguji kemampuan model dalam melakukan prediksi dengan akurasi yang tinggi, dengan melakukan evaluasi performa dari model *Naïve Bayes*, dapat dipastikan apakah model yang dibuat mampu memberikan prediksi yang tepat dan dapat diandalkan. Selanjutnya untuk mengetahui performa dari algoritma *Naïve Bayes*, maka dilakukan pengujian terhadap model yang telah dibuat. Hasil klasifikasi akan divisualisasikan dalam bentuk *Classification Report* dan *Confusion Matrix* yang menghasilkan metrix-metrix seperti *Accuracy*, *Recall*, *Precision*, *F1-Score*, dan *Classification Error*. Gambar 20 menunjukkan hasil *classification report* menggunakan algoritma *Naïve Bayes* menggunakan Bahasa pemrograman *Python*.

	precision	recall	f1-score	support
0	0.77	0.89	0.83	19
1	0.89	0.76	0.82	21
accuracy			0.82	40
macro avg	0.83	0.83	0.82	40
weighted avg	0.83	0.82	0.82	40

Sumber: Hasil Penelitian (2023)

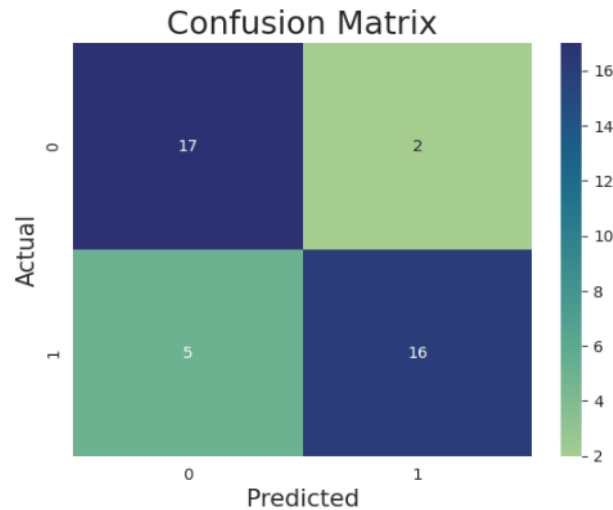
Gambar 20. *Classification Report*

Classification report memberikan evaluasi performa dari model *Naïve Bayes*. Penjelasan masing-masing metrik dalam *classification report* (Nitami & Februariyanti, 2022):

- Precision*: merupakan rasio antara jumlah *true positif* (prediksi benar kelas positif) dengan jumlah total prediksi positif. *Precision* yang lebih tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mengklasifikasikan data positif dengan benar. Pada kelas 0, *precision* sebesar 0.77, artinya sekitar 77% prediksi kelas 0 benar. Pada kelas 1, *precision* sebesar 0.89, artinya sekitar 89% prediksi kelas 1 benar.
- Recall*: merupakan rasio antara jumlah *true positif* dengan jumlah total data yang sebenarnya positif. *Recall* yang lebih tinggi menunjukkan bahwa model memiliki kemampuan yang baik dalam mendeteksi data positif. Pada kelas 0, *recall* sebesar 0.89, artinya sekitar 89% data kelas 0 terdeteksi dengan benar. Pada kelas 1, *recall* sebesar 0.76%, artinya sekitar 76% data kelas 1 terdeteksi dengan benar.
- F1-score*: merupakan rata-rata harmonik dari *precision* dan *recall*. *F1-score* yang lebih tinggi menunjukkan bahwa model memiliki keseimbangan yang baik antara *precision* dan *recall*. Pada kelas 0, *f1-score* sebesar 0.83, dan pada kelas 1, *f1-score* sebesar 0.82.
- Support*: merupakan jumlah data yang termasuk dalam masing-masing kelas.

- e. *Accuracy*: merupakan rasio antara jumlah prediksi benar dengan jumlah total data. Akurasi sebesar 0.82, artinya sekitar 82% data diklasifikasikan dengan benar oleh model.

Pada Gambar 21 hasil klasifikasi dari evaluasi model yang divisualisasikan dalam bentuk *Confusion Matrix* yang menghasilkan metrix-metrix seperti *Accuracy*, *Recall*, *Precision*, *F1-Score*, dan *Classification Error*. Dimana didapatkan jumlah data dari confusion matrix sebanyak 40 data yang didapatkan dari hasil prediksi model *Naïve Bayes* dalam data *testing*.



Sumber: Hasil Penelitian (2023)

Gambar 21. *Confusion Matrix*

Berdasarkan *confusion matrix validation* Gambar 21 dapat diketahui *True Positif* (TP), *False Positif* (FP), *True Negatif* (TN), dan *False Negatif* (FN) antara lain (Pujiyanto & Ristanti, 2019):

- a. *True Positif* (TP) : Didapatkan angka 17 pada posisi (1,1) menunjukkan jumlah prediksi yang benar. Ini berarti ada 17 sampel data yang benar diprediksi sebagai positif oleh model.
- b. *True Negatif* (TN) = Didapatkan angka 16 pada posisi (2,2) menunjukkan jumlah prediksi yang benar. Ini berarti ada 17 sampel data yang benar diprediksi sebagai negatif oleh model.
- c. *False Positif* (FP) = Didapatkan angka 2 pada posisi (2,1) menunjukkan jumlah prediksi yang salah. Ini berarti ada 2 sampel data yang salah diprediksi sebagai positif oleh model.
- d. *False Negatif* (FN) = Dan terakhir didapatkan angka 5 pada posisi (1,2) menunjukkan jumlah prediksi yang salah. Ini berarti ada lima sampel data yang salah diprediksi sebagai negatif oleh model.

3.5 Deployment

Hasil dari analisis data atau model yang dibuat adalah laporan analisis data ulasan yang digunakan sebagai pertimbangan untuk mengembangkan strategi bisnis pada TMLST *Coffee*. Dan, analisis data yang telah diperoleh akan divisualisasikan dalam bentuk jumlah ulasan positif dan negatif pada tampilan *wordcloud*, grafik, *classification report*, dan *confusion matrix*.

4. Kesimpulan

Berdasarkan pada penelitian yang dilakukan, pada model *Naïve Bayes* yang digunakan memiliki tingkat akurasi yang tinggi dalam memprediksi sentimen positif dan negatif, dengan akurasi sebesar 0.82 atau 82%. Selain itu, model ini juga menunjukkan *recall* positif sebesar 0.76 atau 76% dan *recall* negatif sebesar 0.89 atau 89%. Hal ini menunjukkan bahwa model memiliki kemampuan yang baik dalam mengidentifikasi dengan benar ulasan yang sebenarnya memiliki sentimen positif maupun negatif. Tingginya tingkat *recall* negatif menunjukkan bahwa model cenderung lebih baik dalam mendeteksi ulasan yang sebenarnya memiliki sentimen negatif. Dengan begitu, hasil evaluasi model ini memberikan indikasi bahwa *Naïve Bayes* dapat menjadi pilihan yang efektif dalam melakukan analisis sentimen pada data ulasan TMLST *Coffee*.

Daftar Pustaka

- Agustina, N., Citra, D. H., Purnama, W., Nisa, C., & Kurnia, A. R. (2022). Implementasi Algoritma *Naïve Bayes* untuk Analisis Sentimen Ulasan Shopee pada Google Play Store. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 2(1), 47–54. <https://doi.org/10.57152/malcom.v2i1.195>
- Estika, I. Di, Darmawan, I., & Pratiwi, O. N. (2021). *Analisis Sentimen Ulasan Pengguna Untuk Peningkatan Layanan Menggunakan Algoritma Naïve Bayes (Studi Kasus: Bukalapak)*. 8(2), 2735–2745. <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/14676/14453>
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. In *Data Mining: Concepts and Techniques*. Elsevier. <https://doi.org/10.1016/C2009-0-61819-5>
- Kurniawan, D., & Yasir, M. (2022). Optimization Sentimen Analysis using CRISP-DM and *Naïve Bayes* Methods Implemented on Social Media. *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, 6(2), 74. <https://doi.org/10.22373/cj.v6i2.12793>
- Mutawalli, K. A., Muhajirin, A., & Sari, R. (2022). Sistem Penilaian Kinerja Karyawan PT Bank Syariah Indonesia Berbasis Website Menggunakan Algoritma *Naïve Bayes*. *Jurnal Jaring SainTek*, 4(2), 91–104. <https://ejurnal.ubharajaya.ac.id/index.php/jaring-saintek/article/download/1923/1489/5063>
- Nitami, M. T., & Februariyanti, H. (2022). Analisis Sentimen Ulasan Ekspedisi J&T Express Menggunakan Algoritma *Naïve Bayes*. *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, 5(1), 20–29. <https://doi.org/10.36595/misi.v5i1.396>
- Pujianto, U., & Ristanti, P. Y. (2019). Perbandingan Kinerja Metode C4.5 dan *Naïve Bayes* Dalam Klasifikasi Artikel Jurnal PGSD Berdasarkan Mata Pelajaran. *TEKNO*, 29(1), 50. <https://doi.org/10.17977/um034v29i1p50-67>
- Putro, E. N., & Sari, R. (2023). Pemetaan Sasaran Marketing Calon Mahasiswa Baru UBHARA Jaya Menggunakan Algoritma C4.5. *Journal of Students' Research in Computer Science*, 4(1), 63–74. <https://doi.org/10.31599/jsrscs.v4i1.2397>

Widodo, F. (2021). *Membuat Wordcloud Dengan Python*. Accessed: Apr.
<https://sites.unpad.ac.id/widodo/2021/03/01/membuat-wordcloud-dengan-python/>

Yusuf, A. Y. P., & Sari, R. (2022). Implementasi Algoritma *Naïve Bayes* untuk Klasifikasi Pemahaman Program MBKM bagi Mahasiswa. *Journal of Informatic and Information Security*, 3(2), 171–180.