MAKALAH KLASIFIKASI TEKS ULASAN

Disusun untuk Mengikuti Seleksi Datavidia 7.0

Disusun oleh:

Contemporary

Rheco Paradhika Kusuma	10118065
Gabrielle Christy	10818016

INSTITUT TEKNOLOGI BANDUNG

BANDUNG

2021

KLASIFIKASI TEKS ULASAN

I. Pendahuluan

A. Latar Belakang

Pada era revolusi 4.0, data bukanlah sesuatu yang asing dalam kehidupan sehari-hari. Proses pengumpulan data telah dilakukan manusia sejak ribuan tahun yang lalu. Meskipun demikian, data tidak akan berguna apabila tidak ada informasi yang mampu dipetik dari data tersebut. Untuk memperoleh informasi mengenai data, diperlukan proses pengolahan data dan analisis data. Hasil dari analisis data tersebut diharapkan dapat membantu manusia dalam menjawab berbagai persoalan yang ditemui dalam berbagai sektor sehingga manusia dapat meningkatkan kualitas hidupnya. Perkembangan teknologi telah memudahkan manusia melakukan berbagai hal, termasuk analisis data. Seiring dengan berjalannya zaman, muncul berbagai jenis perangkat lunak yang dapat digunakan peneliti untuk memudahkan proses analisis data. Dengan demikian, pada era yang serba canggih ini, analisis data seringkali dikenal dengan istilah *data science. Data Science* merupakan integrasi dari berbagai disiplin ilmu, termasuk pemrograman dan statistik.

Salah satu contoh konkrit pemanfaatan *data science* dapat ditemukan dalam pengklasifikasian teks. Salah satu jenis pengklasifikasian teks ialah analisis sentimen. Dalam analisis sentimen, data yang berisi teks tertentu (misalnya ulasan atau *review*) dari pelanggan (*customer*) diolah sedemikian rupa sehingga mampu dikategorikan menjadi beberapa kelas sesuai dengan sentimen (sebagai contoh, ulasan positif dan ulasan negatif).

Dalam makalah ini, dianalisis data ulasan (*review*) salah satu hotel ternama di Indonesia. Tujuan utama dari penulisan laporan ini ialah untuk menyajikan hasil analisis yang telah dilakukan untuk membantu pemilik perusahaan untuk memprediksi ulasan pelanggan mana yang bersifat positif atau negatif secara lebih efisien. Dengan adanya model yang sesuai, hal ini akan mempermudah para pemangku jabatan atau pemilik perusahaan untuk mengevaluasi performansi usahanya dan dapat dijadikan pertimbangan

dalam mengambil keputusan untuk menentukan strategi yang cocok untuk ke depannya.

B. Rumusan Masalah

Berdasarkan latar belakang yang telah disinggung pada bagian sebelumnya, adapun beberapa rumusan masalah yang ditinjau dalam penulisan makalah ini, antara lain:

1. Bagaimana mengklasifikasikan ulasan yang positif dan yang negatif?
2. Bagaimana performansi dari model klasifikasi yang didapat dan bagaimana hasil prediksinya?

II. Analisis Data

Adapun data yang digunakan merupakan data ulasan (*review*) salah satu hotel ternama di Indonesia. Data *train* terdiri dari 14.856 baris dan 3 kolom, yakni kolom "*review_id"* yang merupakan kode identitas dari pelanggan yang memberikan ulasan, "*review_text"* yang berisi ulasan dari pelanggan, dan "*category",* di mana "*category"* yang bernilai 0 berarti ulasan yang diberikan bersifat negatif dan 1 berarti ulasan yang diberikan bersifat positif. Sedangkan data *test* terdiri dari 3.714 baris dengan 2 kolom, yakni "*review_id"* dan "*review_text"*. Berikut merupakan tahapan yang dilakukan untuk memprediksi "*category*" dari data *test*.

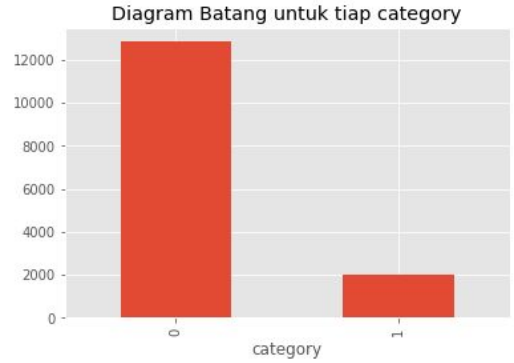
A. *Preprocessing*

Tahapan ini meliputi proses pembersihan data. Pembersihan yang dimaksud ialah menghilangkan data yang kosong dan mengalami duplikat. Dalam data yang dianalisis, tidak ditemukan data yang kosong dan duplikat (dilihat dari "*review_id*" dan "*review_text*" yang bersifat unik). Untuk mempermudah analisis, data "*review_text*" diubah ke format huruf kecil dan mengubah emoji menjadi teks. Selain itu, tanda baca, bilangan, huruf berulang dan karakter spesial (*special characters)* juga dihilangkan. Ada beberapa ulasan yang juga kami hapus dari data karena ulasan yang diberikan kurang relevan, misalnya untuk ulasan yang hanya terdiri dari

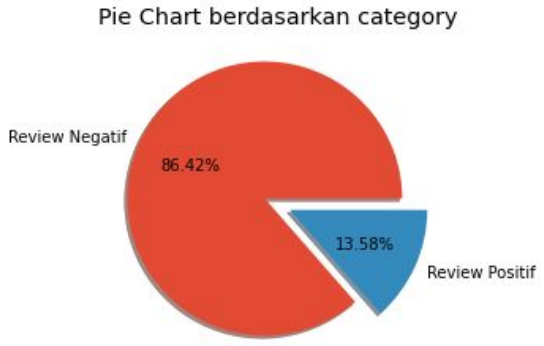
satu atau dua karakter saja. Kemudian, kami juga menggunakan *tokenization* dan *lemmatization* dari *package nltk. Tokenization* merupakan langkah yang digunakan untuk memisahkan sebuah kalimat menjadi bentuk kata-kata. Sedangkan *lemmatization* merupakan suatu proses menemukan lema dalam kata-kata yang digunakan. Terakhir, kami juga melakukan substitusi terhadap beberapa kata, misalnya kata-kata yang disingkat. Sebagai contoh, kami mengubah kata "ga", "ngga", "tdk" agar seragam menjadi "tidak".

B. *Exploratory Data Analysis*

Dalam tahap ini, dilakukan visualisasi dari data yang ada. Berikut merupakan diagram batang dan *pie chart* yang menunjukkan perbandingan antara ulasan yang bersifat positif dan negatif sebelum data dibersihkan.



Gambar 1 Diagram Batang Perbandingan Ulasan Negatif dan Positif



Gambar 2 *Pie Chart* Perbandingan Ulasan Negatif dan Positif

Berdasarkan gambar di atas, dapat dilihat bahwa mayoritas ulasan yang diberikan oleh pelanggan merupakan ulasan yang bersifat buruk atau negatif. Hal ini menandakan bahwa performansi bisnis yang telah dijalankan belum baik.

Selain itu, dilakukan juga visualisasi terhadap kata-kata yang sering digunakan dalam penulisan ulasan.

Gambar 3 Kata-kata yang Sering Digunakan dalam Ulasan Positif sebelum Data Dibersihkan

Gambar 4 Kata-kata yang Sering Digunakan dalam Ulasan Negatif sebelum Data Dibersihkan

Gambar 5 Kata-kata yang Sering Digunakan dalam Ulasan Positif sesudah Data Dibersihkan

Gambar 6 Kata-kata yang Sering Digunakan dalam Ulasan Negatif sesudah Data Dibersihkan

Berdasarkan visualisasi yang telah dilakukan, diperoleh beberapa kata kunci dari ulasan yang bersifat positif, diantaranya ialah bersih, bagus, dan nyaman. Sedangkan, beberapa kata kunci dari ulasan yang bersifat negatif, diantaranya ialah kurang bersih, tidak dingin, dan tidak sesuai.

C. *Feature Engineering*

Setelah itu, dengan bantuan *package sklearn,* dilakukan *countvectorizer* dan *TF-IDF* (*Term Frequency - Inverse Document Frequency*). *Countvectorizer* mempermudah untuk mendeteksi kata-kata yang ada dan sering digunakan oleh pelanggan. Sedangkan *TF-IDF* mempermudah untuk melihat frekuensi penggunaan kata-kata tertentu sekaligus digunakan untuk memberikan bobot pada kata yang sering muncul. Dengan bantuan *TF-IDF,* dilakukan juga normalisasi atas pembobotan kata tersebut.

D. *Modelling*

Sebelum dilakukan pemodelan, kami melakukan pemisahan data menjadi data latih (*train)* dan data evaluasi (*test*) dengan *train test split.* Karena tujuan utama dari pengolahan data ini ialah untuk memprediksi kelas, kami mencoba beberapa model klasifikasi *supervised learning*. Ada beberapa jenis model yang kami gunakan, yakni regresi logistik, SGD (*Stochastic Gradient Descent*) *Classifier*, SVC (*Support Vector Classifier*), dan BERT. Selain itu, kami juga melakukan *hyperparameter tuning* untuk model yang memiliki *macro F1*-s*core* tertinggi.

E. *Validation*

Pada tahap ini, dilakukan evaluasi terhadap model yang telah diperoleh. Karena kami menggunakan model klasifikasi, maka kami menggunakan *classification report* dari *packages sklearn,* terutama *macro F1-score.* Adapun rumusan untuk *macro F1-score* ialah:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

dengan *TP* menyatakan *true positive, FP false positive,* dan *FN false negative. F1-score* ini juga dapat ekivalen dengan rumusan yang dihitung dari *precision* dan *recall.* Dengan demikian, *F1-score* dapat dikatakan sebagai jembatan penghubung antara *precision* dan *recall.*

Berikut merupakan hasil yang didapatkan untuk ketiga model sebelum dilakukan *preprocessing* dan sesudah.

Model	*Preprocess*	*Accuracy*	*F1-Score Macro*	*Accuracy (train)*	*F1-Score Macro (train)*
Regresi Logistik	Sebelum	93.61%	0.84	94.88%	0.87
	Sesudah	93.49%	0.84	95.19%	0.88
SGD *Classifier*	Sebelum	94.28%	0.87	97.09%	0.93
	Sesudah	94.35%	0.87	96.99%	0.93
SVC	Sebelum	94.10%	0.86	98.52%	0.97
	Sesudah	93.99%	0.85	98.36%	0.96

Tabel 1 Performansi Model sebelum dan sesudah *Preprocessing*

Berdasarkan tabel di atas, model yang terbaik ialah model SGD *Classifier* dengan *macro F1-Score 0.87*. Kemudian, kami juga melakukan *hyperparameter tuning* untuk model SGD *Classifier*. Berikut merupakan *classification report* untuk model tersebut.

Model	Normalisasi	*Accuracy*	*F1-Score Macro*	*Accuracy (train)*	*F1-Score Macro (train)*
SGD *Classifier*	Ya	94.37%	0.87	96.90%	0.93
	Tidak	93.94%	0.86	97.61%	0.95

Tabel 2 Performansi Model SGD *Classifier* dengan *Hyperparameter Tuning*

Berdasarkan tabel 2, dengan dilakukannya *hyperparameter tuning,* performansi model tidak berubah secara signifikan berdasarkan *macro F1-Score.* Dengan demikian, dapat dikatakan bahwa sejauh ini, model SGD *Classifier* dengan *hyperparameter tuning* dan normalisasi merupakan model yang terbaik. Terakhir, kami melakukan modelling dengan menggunakan BERT dengan *epoch* 2 dan *batch size* 64.

Model	*Preprocess*	*F1-Score Macro*
BERT	Sesudah	0.90

Tabel 3 Performansi Model BERT

Apabila dibandingkan dengan semua model yang telah dicoba. Maka, model yang terbaik ialah model BERT dengan *preprocess*, yakni dengan *macro F1-Score* 0.90.

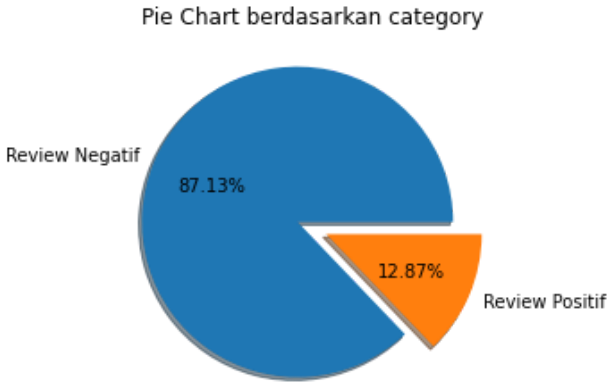
III. Pembahasan

Berdasarkan model yang didapatkan, kami melakukan prediksi untuk data *test (test.csv)* yang disediakan dan berikut merupakan hasilnya untuk 5 data pertama.

	review_id	review_text	category
0	7302180ac7160f04a405d8aa7bd6beb8	kasur rusak punggung saya sakit kasurnya tidak...	0
1	3184c670e170f565c7522eb76a320ba1	gerah	0
2	e33abf6bb5d5a9a77c339043b1725dc8	tempat ridur atas bawah ac nya pararel tidak a...	0
3	061d388950340070a6ac03fabc9027b0	ac kurang dingin tidak ada snack sperti di foto	0
4	e62d87c348674b6c06856964f3ab16b8	lumayan untuk guest house dengan standart harg...	1

Gambar 7 Hasil Prediksi Data *Test*

Jika dilakukan visualisasi, berikut merupakan *pie chart* perbandingan ulasan positif dan negatif yang diberikan oleh pelanggan pada data *test.*



Gambar 8 *Pie Chart* Hasil Prediksi Data *Test*

Berdasarkan grafik di atas, sama halnya dengan data *train*, diperoleh bahwa mayoritas ulasan yang diberikan oleh pelanggan merupakan ulasan yang bersifat buruk atau negatif. Hal ini menjadi "sinyal buruk" bagi bisnis yang dijalankan karena memiliki reputasi yang buruk.

Dalam melakukan pemodelan, kami beberapa kali melakukan pembersihan terhadap data yang dimiliki. Namun, semakin "bersih" data *train,* didapatkan performansi model yang semakin buruk. Hal ini kemungkinan dikarenakan data tidak sepenuhnya bersih.

Selain itu, sebenarnya kami telah mencoba melakukan pemodelan terhadap beberapa model *deep learning.* Akan tetapi, hasil yang didapatkan belum optimal dan kami mengalami kendala dalam melakukan proses *training* dari model *deep learning* karena membutuhkan waktu yang sangat lama dan laptop yang memadai.

IV. Simpulan dan saran

Untuk mengklasifikasikan ulasan yang positif dan yang negatif, digunakan beberapa model, yakni regresi logistik, SGD (*Stochastic Gradient Descent*) *Classifier*, SVC (*Support Vector Classifier*), dan BERT. Berdasarkan keempat model tersebut, model yang memberikan performansi terbaik ialah model *BERT* dengan *macro F1-score* 0.90. Hasil prediksi dari model yang didapatkan terdiri dari 87.13% ulasan negatif dan 12.87% ulasan positif. Hal ini menandakan bahwa perusahaan yang dianalisis memiliki reputasi yang kurang baik dan sebaiknya ditingkatkan.

Model yang didapatkan pada laporan ini merupakan model yang umum dipakai dalam pengklasifikasian teks. Disarankan untuk mencoba model-model *deep learning* yang lebih kompleks dan sesuai untuk kasus pengklasifikasian teks, seperti LSTM. Kemudian, pada tahap *preprocessing* dan *feature engineering,* akan lebih baik lagi jika dilakukan pembobotan terhadap kata-kata yang signifikan dan relevan saja (misalnya, kata-kata yang bukan merupakan konjungsi), menggunakan ngram, *stopword* dan *stemmer* dari *package* Sastrawi sehingga mungkin model dan hasil prediksi yang diberikan akan menjadi lebih baik.

DAFTAR PUSTAKA

Koenig, Rachel. (2019, 29 Juli). *NLP for Beginners: Cleaning & Preprocessing Text Data*. Diakses dari https://towardsdatascience.com/nlp-for-beginners-cleaning-preprocessing-text-data-ae8e306bef0f.

Monsters, Data. (2018, 16 Oktober). *Text Preprocessing in Python: Steps, Tools, and Examples*. Diakses dari https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908.

Scikit Learn. *Feature Extraction Text: TF-IDF Transformer*. Diakses dari https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html.

Zhang, A., et al. (2021, 6 Januari). *Dive into Deep Learning* (*release* 0.16.0). Diakses dari d2l.ai.