REVIEW SENTIMEN ANALISIS APLIKASI SOSIAL MEDIA DI GOOGLE PLAYSTORE MENGGUNAKAN METODE LOGISTIC REGRESSION

Edward Darmaja[1], Viny Christanti Mawardi[2], Novario Jaya Perdana[3]

[1] Jurusan Teknik Informatika, Universitas Tarumanagara
Email: Edward.535180043@stu.untar.ac.id
[2] Jurusan Teknik Informatika, Universitas Tarumanagara
Email: vinym@fti.untar.ac.id
[3] Jurusan Teknik Informatika, Universitas Tarumanagara
Email: novariojp@fti.untar.ac.id

ABSTRACT

Everyone has their own nature and way of thinking, which makes their style of conveying what they are thinking is different from one to another. Not only that, what they are conveying can have a different meanings according to the person who listens to it and which perspective they take it into. From this, the reviews given have different meanings for each person which other people can easily understand the meaning behind it. Although humans can distinguish whether the review given by other people has a positive or negative meaning, machine cannot understand this without being given instructions first, in order for machine to find out the meaning in the text, Therefore, this research was conducted. The purpose of designing this system is to make it easier to analyze a collection of reviews on a social media application from google playstore without the need to see all the review given by other users. This system is also designed to fulfill the purpose of evaluating using the TF-IDF method and the Logistic Regression method on the program to be created. The intended evaluation is from the level of accuracy obtained from the use of the model formed using the method chosen and how many predictions have the correct value based on the result of Confusion Matrix from the program that has been designed . From the results obtained, it can be concluded whether the social media application has a good reputation from its users or not.

Keywords: *Natural Language Processing (NLP); Sentiment Analisist; TF-IDF; Logistic Regression*

ABSTRAK

Setiap orang memiliki sifat serta cara pemikirannya sendiri, yang membuat gaya cara mereka saat menyampaikan apa yang sedang mereka pikirkan tidak sama satu dengan lainnya. Tidak hanya itu apa yang mereka sampaikan tersebut dapat memiliki arti serta makna yang berbeda bagi sisi setiap orang yang menyimaknya. Dari hal tersebut ulasan yang diberikan setiap orang memiliki makna yang berbeda-beda antara satu dengan yang lainnya yang dimana manusia dapat dengam mudah memahami hal tersebut. Walaupun manusia dapat membedakan apakah teks ulasan yang dibaca tersebut terkesan positif maupun negatif, mesin belum dapat memahami makna yang tersirat dari teks tersebut tanpa instruksi yang diberikan, maka penelitian agar mesin dapat mengetahui makna yang ada pada teks dilakukan. Adapun tujuan dari perancangan sistem ini yaitu untuk memudahkan dalam melakukan analisis dari kumpulan ulasan pada suatu aplikasi sosial media tanpa perlu melihat satu-persatu setiap ulasan yang diberikan pengguna lainnya. Sistem ini juga dirancang untuk memenuhi tujuan untuk evaluasi menggunakan metode TF-IDF dan metode Logistic Regression pada program yang akan dibuat. Evaluasi yang dimaksudkan yaitu tingkat akurasi yang diperoleh hasil dari penggunakan model yang dibentuk menggunakan metode yang dipilih pada pembuatan serta berapa banyak hasil prediksi yang memiliki nilai benar berdasarkan Confusion Matrix yang dihasilkan pada Hasil program yang telah dirancang. Dari hasil yang diperoleh tersebut dapat disimpulkan apakah aplikasi sosial media tersebut memiliki reputasi yang baik dari para penggunanya atau tidak.

Kata Kunci: Natural Language Processing (NLP); Sentimen Analisis; TF-IDF; Logistic Regression

1. PENDAHULUAN

Latar Belakang

Setiap orang memiliki sifat serta cara pemikirannya sendiri, yang dimana membuat setiap orang memiliki gaya cara mereka tersendiri saat menyampaikan apa yang sedang mereka pikirkan, tidak hanya itu apa yang mereka sampaikan tersebut dapat memiliki arti serta makna yang berbeda bagi sisi setiap orang yang menyimaknya.

Walaupun teks yang dibaca terkesan positif, belum tentu kenyataanya itu berisikan hal positif, sebab sesuai cara pandang orang tersebut saat memahami makna yang terdapat pada teks maka akan berbeda juga kesan yang diberikan. Seperti saat seseorang memberikan tanggapan atau ulasan pada suatu produk, mereka akan menuliskan apa yang mereka pikirkan setelah menggunakan produk tersebut sebagai bahan referensi untuk orang lain yang akan menggunakannya kedepan.

Ketika mengulas suatu karya, pengulas harus bersikap kritis agar hasil ulasannya dapat memberikan kontribusi demi kemajuan karya tersebut. Tentunya di dalam kehidupan sehari-hari kerap menemukan berbagai jenis ulasan terkait suatu film, buku, lagu, hingga karya sastra dan seni lainnya di dalam surat kabar atau media lainnya, yang mana hal tersebut disebut sebagai teks ulasan.

Berdasarkan teks ulasan yang diberikan tersebut, pembuat dapat mengetahui apa yang sedang dipikirkan oleh penggunanya dan memprediksikan apa yang diinginkan serta mengetahui mengapa produk yang mereka buat tidak diminati ataupun mengapa produk tersebut banyak digunakan. Dari hal tersebut pembuat dapat menggunakannya sebagai referensi untuk kedepannya untuk meningkatkan kualitas produk yang mereka buat agar dapat membuat para penggunanya tetap senang dan setia untuk menggunakan produk yang mereka buat tersebut. Walaupun manusia dapat membedakan apakah teks ulasan yang dibaca tersebut positif maupun negatif, mesin belum dapat memahami makna yang tersirat dari teks tersebut tanpa instruksi yang diberikan, maka penelitian agar mesin dapat mengetahui makna yang ada pada teks dilakukan.

Dari Machine Learning akan menjawab pertanyaan tentang bagaimana suatu program pada komputer atau mesin dapat meningkatkan performa yang diberikan berdasarkan dari pengalaman tugas yang telah diberikan sebelumnya (Tom Mitchell, 1997, p.17). Algoritma yang digunakan pada Machine Learning juga sudah terbukti memiliki nilai praktik yang baik pada berbagai bidang, salah satunya seperti penerapan pada Data Mining, pada kasus ini terdapat database berukuran besar yang berisikan banyak data yang tidak teratur, dimana dengan pemanfaatan Machine Learning data yang berharga atau penting tersebut dapat ditemukan secara otomatis (Tom Mitchell, 1997, p.17).

Natural Language Processing merupakan bagian dari Artificial Intelligence (AI) yang memberikan mesin kemampuan untuk membaca, memahami, dan menafsirkan bahasa manusia. Hal tersebut membantu mesin dalam mengukur nilai sentimen dan mendeterminasikan bagian dari bahasa manusia yang mana merupakan hal yang penting. Untuk mesin hal ini merupakan sesuatu yang sulit untuk dilakukan dikarnakan banyaknya bentuk data yang tidak terstruktur, kurangnya peraturan yang formal dan tidak ditemukannya maksud atau tujuan sebenarnya.

Sentiment Analysis atau biasa disebut juga dengan opinion mining, merupakan suatu teknik pada Natural Language Processing yang digunakan untuk mendeterminasi apakah data tersebut positif, ataupun negatif. Sentiment Analysis sering digunakan pada data berbentuk teks untuk membantu dalam memonitori sentimen ulasan yang diberikan oleh pengguna dan memahami apa yang pengguna butuhkan.

Berikut merupakan beberapa penelitian sejenis yang pernah dikerjakan sebelumnya. Pada suatu penelitian yang dilakukan oleh Pang, Lee, dan Vaithyanathan, mengusulkan penggunaan *sentiment classification* menggunakan *machine learning* pada movie review dataset. Penganalisaan tersebut menggunakan metode *Naïve Bayes*, *Max Entropy* dan *Support Vector Machine* (SVM) model pada data unigram dan bigram. Pada percobaan tersebut hasil yang diperoleh dari penggunaan SVM bersama unigram dengan *feature extraction* menghasilkan akurasi yang sangat baik yaitu sebesar 82.9% (Nguyen, Heidi, Aravind Veluchamy, Mamadou Diop, & Rashed Iqbal, 2018, p.7).

Pada suatu penelitian yang dilakukan oleh Warnia Nengsih, M.Mahrus Zein, dan Nazifa Hayati, melakukan analisis sentimen terhadap ulasan hotel. Pada penelitian tersebut menggunakan metode *Random Forest* (RF) sebagai metode pengklasifikasian ulasan. Hasil yang diperoleh yaitu untuk tingkat akurasi prediksi sebesar 90% yang dapat dibilang sangat tinggi dan baik serta untuk ulasan positif sebesar 68% dan negatif sebesar 32% (Nengsih, Warnia, M. Mahrus Zein, & Nazifa Hayati, 2021, p.41-48).
Pada suatu penelitian yang dilakukan oleh Emilie Coyne, dan Jim Smit, melakukan analisis sentimen terhadap kumpulan ulasan produk dari Amazon. Pada penelitian tersebut menggunakan perbandingan dari beberapa metode seperti *Linear SVM*, *Multinominal NB*, dan *LSTM network* sebagai pengklasifikasian ulasannya. Dari hasil prediksi yang diberikan, dari ketiga metode tersebut memperoleh nilai akurasi prediksi yang sangat besar yaitu semuanya diatas 90% dan itu terbukti sangat akurat dan baik untuk digunakan (Güner, Levent, Emilie Coyne, & Jim Smit, 2019, p.3-7).

Rumusan Masalah

Adapun rumusan masalah yang menjadi acuan dari latar belakang sebelumnya yaitu, bagaimana cara agar mesin dapat mengenali teks ulasan, dan bagaimana teks ulasan tersebut akan diproses agar dapat diketahui masuk ke kategori mana ulasan yang telah diproses tersebut apakah postif, ataupun negatif.
Untuk dapat melakukan hal tersebut, dirancang suatu program aplikasi ini. Untuk pembuatannya akan menggunakan bahasa Python serta dibantu dengan beberapa library lain yang dibutuhkan seperti Tkinter sebagai pembantu pembuatan GUI-nya (Graphical User Interface), atau dapat dikatakan sebagai User Interface.
Sebagai input yang akan digunakan yaitu file excel dengan format ".xlxs" yang berisikan ulasan suatu aplikasi sosial media. Sedangkan untuk outputnya akan berbentuk keterangan tentang berapa banyak data yang diolah, jumlah ulasan positif, dan ulasan negatif serta representasinya dalam bentuk Pie Chart dan Bar Chart.

2. METODE PENELITIAN

Terdapat dua buah metode yang digunakan pada penelitian ini untuk membuat rancangan program yang akan dibuat, diantaranya yaitu TF-IDF, dan Logistic Regression. Untuk penjelasannya dapat dilihat dibawah ini.

TF-IDF

TF-IDF atau dapat dikatakan "*Term Frequency – Inverse Document Frequency*" merupakan suatu teknik yang digunakan untuk mengukur kata pada dokumen dengan menghitung bobot setiap kata pada suatu dokumen. Teknik ini banyak digunakan pada *Information Retrieval* dan *Text Mining*. Terdapat beberapa tahap yang akan dilakukan untuk melakukan TF-IDF ini, yang diantaranya yaitu :

1. *Term Frequency* (TF)
   Merupakan tahap untuk menghitung seberapa sering kemunculan suatu kata pada dokumen. Untuk seberapa banyak frekuensi kata tersebut tergantung dari seberapa panjang teks atau dokumen yang dimiliki.
   $$tf(t,d) = count\ of\ t\ in\ documents \quad (1)$$
2. *Document Frequency* (DF)
   Hampir sama seperti pada tahap sebelumnya, perbedaanya yaitu DF akan menghitung kemunculan setiap kata sebelumnya berdasarkan seberapa banyak kemunculan kata tersebut pada suatu dokumen.
   $$df(t) = occurrence\ of\ t\ in\ documents \quad (2)$$

3. *Inverse Document Frequency* (IDF)
   Tahap terakhir ini merupakan suatu tahap untuk menghitung bobot dari suatu teks atau dokumen berdasarkan hasil perolehan TF dan DF pada langkah sebelumnya. Terdapat beberapa rumus yang perlu dilakukan untuk memperoleh bobot tersebut yang diantaranya adalah :

$$idf(t) = Log(\frac{d}{DF}) \quad (3)$$

Dimana :
t = term (kata)
d = document (Jumlah dokumen)
DF = Document Frequency (frekuensi dokumen berdasarkan kata)

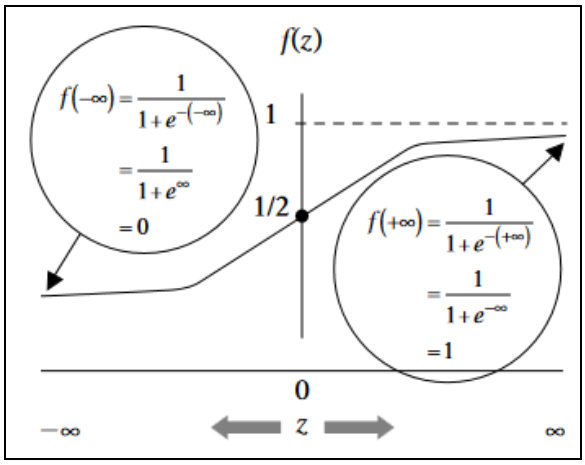
$$W = TF * (IDF + 1) \quad (4)$$

Dimana :
W = weight (bobot kata)
TF = Term Frequency (frekuensi kemunculan kata)
IDF = Inverse Document Frequency

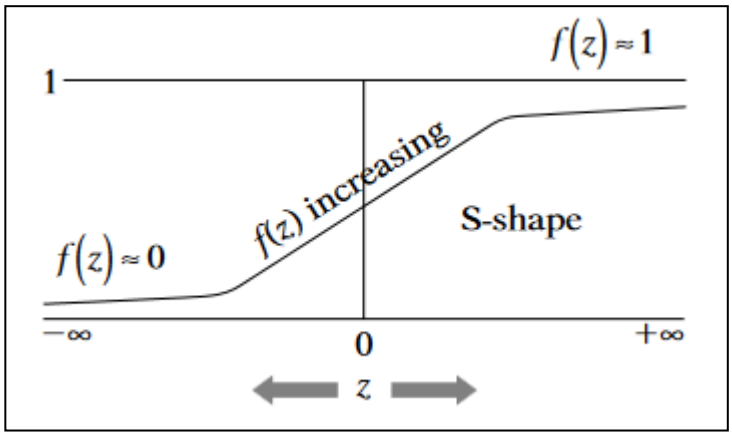
Logistic Regression

Logistic Regression merupakan suatu model matematika yang cara pendekatannya dapat digambarkan dengan hubungan dari beberapa variabel X hingga variabel D yang bersifat *dichotomous dependent variable* (David G. Kleinbaum & Mitchel Klein, 1994, p.5-6). *Dichotomous dependent variable* artinya yaitu suatu variable hanya memiliki dua buah plilihan, yang dimana harus memilih dari salah satu pilihan ketika sedang melakukan observasi atau pengukuran.

Logistic Regression popular untuk digunakan disebabkan dari hasil *logistic function f(z)* yang dimana hasil yang diberikan diantara 0 dan 1. Model tersebut di desain untuk menjelaskan tentang probabilitas yang dimana selalu diantara nilai 0 dan 1. Salah satu contohnya yaitu probabilitas dari suatu individual terjangkit suatu penyakit (David G. Kleinbaum & Mitchel Klein, 1994, p.5-6). Pada "Gambar 1. Logistic Function" dibawah, dapat dilihat bahwa hasil yang akan diberikan yaitu antara nilai 0 dan 1, walaupun nilai masukkan untuk variabel *z* itu sendiri memiliki masukkan nilai yang berbeda-beda (David G. Kleinbaum & Mitchel Klein, 1994, p.5-6).



Gambar 1. Logistic Function
Sumber Gambar: David G. Kleinbaum and Mitchel Klein, Logistic Regression: A Self-Learning Text, 2nd Edition, (Springer, New York, NY, (1994)), h.6.



Gambar 2. Slope Logistic Function
Sumber Gambar: David G. Kleinbaum and Mitchel Klein, Logistic Regression: A Self-Learning Text, 2nd Edition, (Springer, New York, NY, (1994)), h.6.

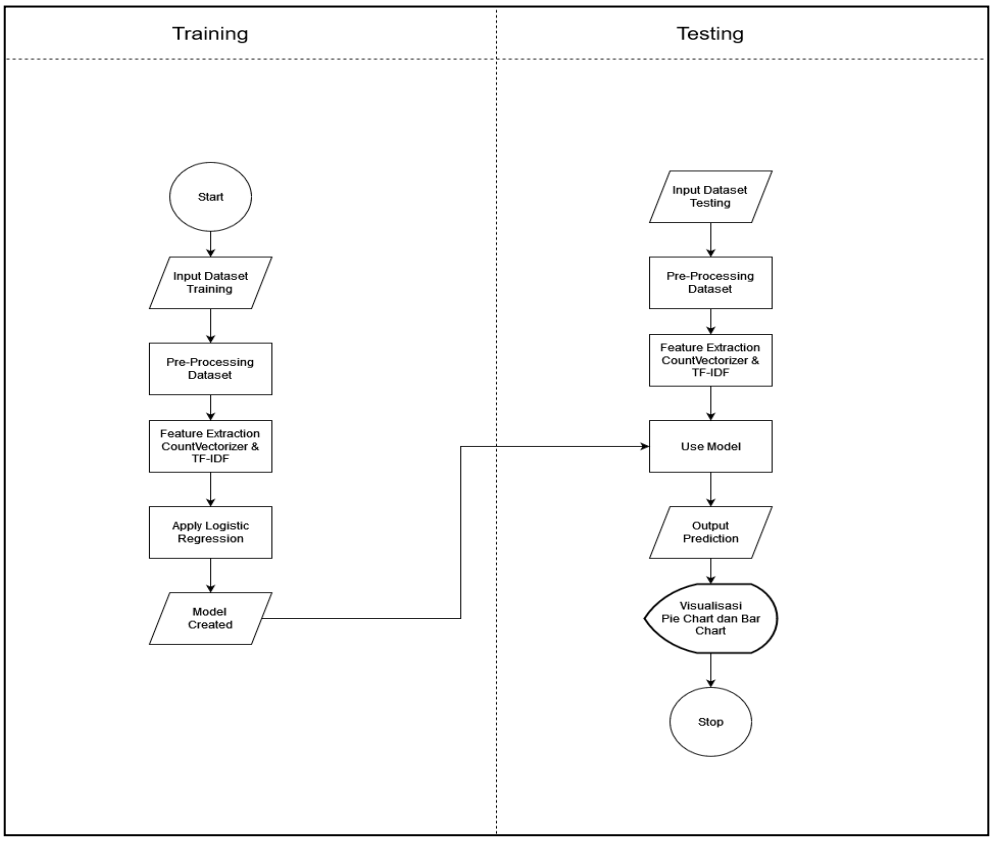
Pada "Gambar 2. Slope Logistic Function" diatas, dapat dilihat bahwa semakin tinggi atau besar nilai yang dimasukkan (z), maka hasil yang akan diperoleh adalah 1 sedangkan jika nilai yang dimasukkan (z) kecil, maka hasil yang akan diperoleh adalah 0. Dari hal tersebut, *Logistic Model* kemudian dapat dipastikan bahwa apapun estimasi dari hasil yang diperoleh akan selalu menghasilkan angka diantara 0 dan 1. Maka daripada itu hasil yang diberikan dari *Logistic Model* tidak akan lebih dari 1 ataupun dibawah 0. Untuk model lain tidak mungkin dapat diperoleh hal tersebut, itu lah sebabnya mengapa *Logistic Model* ini sering digunakan ketika akan melakukan probabilitas pada suatu data (David G. Kleinbaum & Mitchel Klein, 1994, p.5-6).
Untuk sistem yang akan dirancang menggunakan metode ini untuk membagi ulasan kedalam dua kategori yaitu apakah ulasan masuk kedalam kategori positif ataupun negatif untuk netral itu sendiri tidak akan digunakan sebab pada *Logistic Regression* hasil keluaran yang diperoleh yaitu 0 dan 1 untuk dua kategori positif dan negatif maka netral tidak akan digunakan. Untuk pengkategorian tersebut dapat diperoleh dengan hasil output model dari *Sigmoid Function* atau *Logistic Function* yang digunakan pada *Logistic Regression*.
Setelah diperoleh hasil bobot suatu ulasan menggunakan metode TF-IDF, hasil bobot tersebut kemudian akan digunakan pada *Sigmoid Function*. *Sigmoid Function* merupakan bagian penting untuk pembuatan model *Logistic Regression* yang dimana untuk definisinya sebagai berikut :

$$sigmoid(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

Keterangan :
z = data input
Setelah fungsi tersebut digunakan, hasilnya kemudian akan dibandingkan untuk penentuan kategori dengan kriteria pada umunya sebagai berikut :
1. Jika Hasil > 0.5 maka hasil prediksinya adalah 1.
2. Jika Hasil < 0.5 maka hasil prediksinya adalah 0.

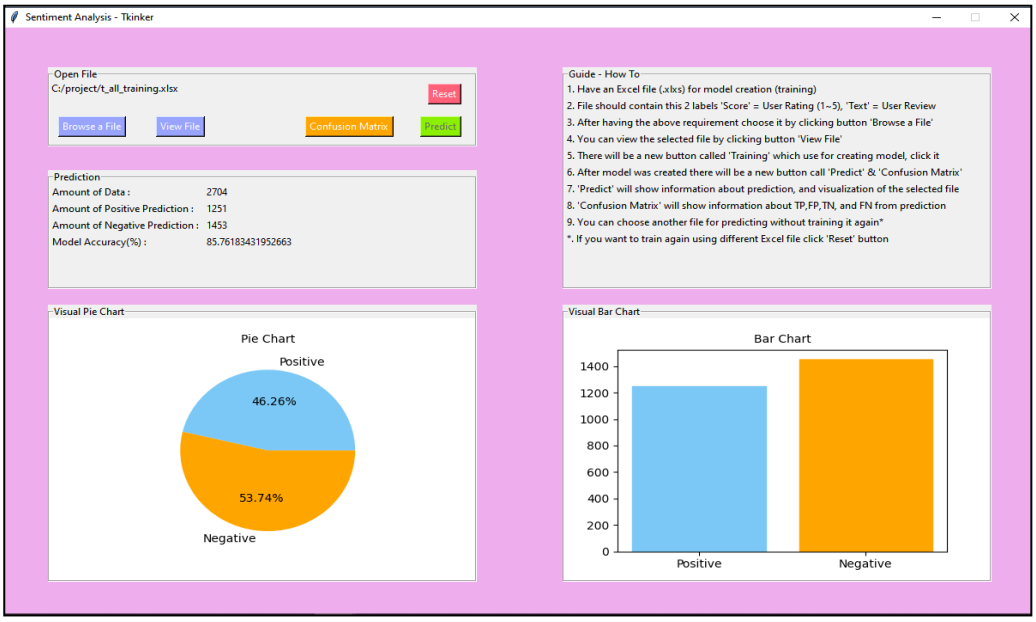
3. HASIL DAN PEMBAHASAN

Untuk hasil dari penelitian yang dilakukan, yaitu dihasilkan suatu aplikasi yang dimana nantinya akan digunakan sebagai media untuk analisis sentiment ulasan suatu aplikasi sosial media dari Google Playstore. Untuk proses kerja dari rancangan yang dihasilkan dapat dilihat pada "Gambar 3.Flowchart Rancangan" dibawah ini.



Gambar 3. Flowchart Rancangan

Pembuatan dilakukan pertama pada media *Google Colab*. Disana akan dilakukan pembentukan serta penyusunan kode, jika sudah tidak ditemukan error dan memiliki hasil yang sesuai, maka akan dilanjutkan menggunakan Visual Studio Code untuk pembuatan tampilan serta pemasangan kode yang telah dibuat. Untuk hasil akhirnya dapat dilihat pada "Gambar 4. Hasil Aplikasi" dibawah ini.



Gambar 4. Hasil Aplikasi

Dari aplikasi tersebut dapat dilihat informasi berupa berapa banyak data yang digunakan, jumlah ulasan positif dan negatif hasil prediksi, akurasi yang diperoleh, serta hasil visualisasinya dalam bentuk diagram Pie dan Bar. Jika ingin melihat file yang digunakan dalam bentuk tabel dan confusion matrix yang dihasilkan dapat menggunakan tombol yang tersedia ("View File", dan "Confusion Matrx").

4. KESIMPULAN DAN SARAN

Dari penggunaan metode TF-IDF dan Logistic Regression dapat diperoleh suatu hasil dimana ulasan tersebut positif ataupun negatif, yang kemudian akan diterapkan pada keseluruhan ulasan yang telah dikumpulkan. Hasil tersebut kemudian dapat ditarik suatu kesimpulan, dimana informasi yang diperoleh dari hasil analisis sentimen ini dapat disimpulkan apakah suatu aplikasi sosial media yang dianalisis tersebut memiliki reputasi yang baik ataupun buruk berdasarkan hasil prediksi penggunaan metode pada ulasan yang diberikan oleh para penggunanya.
Penulis berharap dari penelitian yang dilakukan ini hasil rancangan tersebut dapat berguna untuk memudahkan bagi para pengguna yang ingin mengetahui reputasi suatu aplikasi sosial media tanpa harus membaca semua ulasan yang ada. Kemudian juga dari penelitian ini merupakan wujud dari evaluasi penggunaan metode yang dipakai, harapan dari penulis yaitu kedepannya terdapat penelitian dengan tema yang sejenis tetapi menggunakan metode yang berbeda untuk melihat perbandingan antara metode yang digunakan oleh penulis dan metode lainnya.

Ucapan Terima Kasih (*Acknowledgement*)
Kelancaran melakukan penelitian rancangan ini tidak lepas dari bantuan berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karna itu penulis mengucapkan terima kasih kepada segenap pihak yang membantu ibu Viny Christanti Mawardi, S.Kom., M.Kom. selaku pembimbing utama dalam penelitian yang dilakukan, bapak Novario Jaya Perdana, S.Kom., M.T. selaku pembimbing pendamping dalam penelitian yang dilakukan dan Fakultasi Teknologi Informasi (FTI), selaku penyedia dana untuk melakukan penelitian yang dilakukan. Demikian ucapan terima kasih yang dapat penulis sampaikan, mohon maaf jika ada kata yang salah atau kurang. Penulis mengucapkan terima kasih.

REFERENSI

David G. Kleinbaum & Mitchel Klein, (1994), "Logistic Regression: A Self-Learning Text, 2nd Edition", Springer, New York, NY, h. 5 - 6.

Güner, Levent, Emilie Coyne, & Jim Smit. (2019). "Sentiment analysis for amazon. com reviews.". Big Data in Media Technology (DM2583) KTH Royal Institute of Technology, Stockholm, h.3 – 7.

Tom Mitchell, (1997), "Machine Learning", McGraw-Hill Science/Engineering/Math, h. 17.

Nengsih, Warnia, M. Mahrus Zein, & Nazifa Hayati. (2021). "Coarse-Grained Sentiment Analysis Berbasis Natural Language Processing–Ulasan Hotel.". Jurnal Nasional Teknik Elektro dan Teknologi Informasi 10.1:, h.41-48.

Nguyen, Heidi, Aravind Veluchamy, Mamadou Diop, & Rashed Iqbal. (2018). "Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches". SMU Data Science Review 1, no. 4, h.7.

(*halaman kosong*)