Sentimen Analisis pada Ulasan *Google Play Store* Menggunakan Metode *Naïve Bayes*

Edyt Daryfayi Putra Daulay[1], Ibnu Asror[2]

[1,2,]Fakultas Informatika, Universitas Telkom, Bandung
[1]chrisstone@students.telkomuniversity.ac.id, [2]iasror@telkomuniversity.ac.id,

Abstrak

Ulasan atau biasa disebut review merupakan salah satu fitur penting dari aplikasi yang ada pada Google Play Store. Fitur ini dapat digunakan oleh para pengguna untuk menilai serta memberikan pendapat berupa teks ulasan terhadap aplikasi yang digunakan. Namun untuk beberapa kasus, terdapat teks ulasan yang tidak sesuai dengan nilai atau *rating* yang diberikan. Contohnya jika pengguna memberikan *rating* bintang 5 namun memberikan teks ulasan yang bernada negatif. Penelitian ini membangun mesin klasifikasi yang dikhususkan untuk membandingkan teks ulasan yang diberikan oleh pengguna dengan *rating* yang diberikan. Metode yang digunakan yaitu *Naïve Bayes*, khususnya *Multinomial Naïve Bayes* untuk memudahkan proses klasifikasi karena metode ini dikhususkan untuk klasifikasi teks. Dari hasil penelitian, didapat akurasi setinggi 78,9%, untuk ulasan yang memiliki *rating* bintang 5 dan bintang 1. Tetapi akurasi menurun mencapai 73,7% untuk semua *rating* ulasan terkecuali bintang 3.

Kata kunci : klasifikasi, sentimen analisis, ulasan aplikasi, *naïve bayes, multinomial Naïve Bayes*

Abstract

Review is one of the most important feature of application on Google Play Store. Review can be used by user to rate and share their experince with the application with text review. But for some cases, there are some review that don't align with the star rating that was given from user. For example, if user gave a 5 star rating but their text review are written as negative experience. This research builds a classifier that can be used to compare the text review with the star rating. The method that used in this research is Naïve Bayes, especially Multinomial Naïve Bayes to ease the classification process because Multinomial Naïve Bayes is specialized in text classification. From this research, the highest accuracy that produced is 78,9% for review with 5 star and 1 star review, and the lowest with 73,7% for all rating except 3 star review.

Keywords: classification, sentiment analysis, application review, *naïve bayes, multinomial naïve bayes*

1. Pendahuluan

Latar Belakang

Ulasan suatu produk yang tersebar di berbagai media penting untuk dicermati. Penelitian yang dilakukan pada awal 2013 menunjukkan bahwa ulasan produk yang tersebar di sosial media telah mempengaruhi 90% keputusan seseorang terhadap pembelian produk tersebut[1]. Khusus untuk aplikasi, ulasan terhadap aplikasi tercantum pada *Google Play Store*. Pada halaman *Google Play Store*, ulasan yang diberikan yaitu berupa rating bintang dari satu sampai lima ditambah dengan ulasan teks. Aplikasi yang memiliki *rating* bintang tinggi akan direkomendasikan oleh *Google* dan muncul pada halaman depan atau muncul pada aplikasi terbaik jika diurutkan melalui *rating*. Calon pengguna pada umumnya ingin mengetahui pendapat atau pengalaman dari pengguna lain terkait dengan aplikasi yang akan digunakan dan akan mencari aplikasi berdasarkan *rating* terbaik[10]. Namun terdapat beberapa kejanggalan pada ulasan di *Google Play Store* dimana seorang pengguna memberikan rating kecil seperti satu atau dua bintang, namun teks ulasan yang diberikan merupakan suatu *feedback* positif atau pengguna memberi rating tinggi tetapi ulasan teks yang diberikan memiliki nilai negatif. Akibatnya, *Google* tidak dapat membedakan ulasan positif dengan ulasan negatif melalui teks ulasan yang diberikan pengguna dan dapat mempengaruhi *rating* yang sebenarnya dari aplikasi tersebut.

Untuk mengatasi masalah ini, dibangun sistem klasifikasi untuk mendapatkan sentimen dari ulasan yang diberikan dan membandingkan sentimen tersebut dengan *rating* bintang yang diberikan dengan metode *Naïve Bayes*, khususnya *Multinomial Naïve Bayes* (MNB). MNB merupakan metode *supervised learning* yang menggunakan probabilitas. *Multinomial Naïve Bayes* lebih difokuskan untuk klasifikasi teks [2]. Alasan penulis untuk menggunakan metode MNB yaitu karena MNB memiliki akurasi terbaik untuk mengklasifikasi teks dibandingkan dengan metode lainnya dengan akurasi sebesar 80,2%[3].

Topik dan Batasannya

Topik yang dibahas pada penlitian ini yaitu membangun sistem *classifier* dengan metode MNB dan menganalisis akurasi yang didapat dari metode MNB. Teks ulasan akan diolah pada *preprocessing* terlebih dahulu untuk mendapatkan kata yang dianggap memiliki nilai sentimen positif atau negatif. Kemudian semua kata yang didapat dibandingkan jumlahnya untuk mendapat sentimen umum dari ulasan yang diberikan. Setelah didapat sentimen dari ulasan tersebut akan dibandingkan dengan *rating* bintang yang diberikan. Data ulasan yang digunakan yaitu diambil langsung dari 10 aplikasi yang ada pada *Google Play Store* sebanyak 1500 teks ulasan terkecuali ulasan yang memiliki *rating* bintang 3 berbahasa Indonesia dan dibagi menjadi 150 ulasan untuk masing-masing aplikasi. Hal ini dilakukan karena akurasi yang dihasilkan akan lebih kecil apabila dokumen ulasan yang digunakan terlalu banyak [4].

Tujuan

Tujuan yang dicapai dari penlitian ini adalah mengklasifikasi teks ulasan untuk mendapat sentimen yang diberikan dari ulasan tersebut dan menganalisis performa sistem klasifikasi sentimen dengan metode MNB.

Organisasi Tulisan

Adapun untuk organisasi tulisan dari jurnal yang dikerjakan dibagi menjadi pendahuluan, studi terkait, sistem yang dibangun, evaluasi, dan kesimpulan. Untuk pendahuluan berisikan latar belakang, topik dan bahasan, tujuan, dan organisasi tulisan. Untuk studi terkait berisikan studi literatur, yaitu studi-studi yang terkait dengan jurnal yang dikerjakan. Untuk sistem yang dibangun berisikan gambaran umum sistem, *dataset*, *preprocessing*, dan skenario pengujian. Evaluasi berisikan penjelasan metrik yang digunakan untuk evaluasi, hasil pengujian, dan analisis hasil pengujian. Untuk kesimpulan berisikan kesimpulan yang didapat dari jurnal yang dikerjakan dan saran untuk pengembangan studi dari jurnal yang dikerjakan.

2. Studi Terkait

Studi Literatur

Dari penelitian yang dilakukan oleh Pang (2002) tentang klasifikasi sentimen menggunakan teknik Machine Learning. Pada penelitian ini dilakukan klasifikasi menggunkan tiga metode, yaitu Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), dan Maximum Entropy (ME) untuk mengklasifikasi ulasan film ke dalam kelas ulasan positif dan ulasan negatif. Hasil dari penelitian tersebut menunjukkan bahwa akurasi dari metode NBC memiliki akurasi yang tinggi, yaitu sebesar 80,2% [3].

Dari penelitian yang dilakukan oleh Song (2017) tentang klasifikasi tweet pada Twitter berbasis sentimen analisis. Pada penelitian ini dilakukan klasifikasi menggunakan dua metode, yaitu *Multinomial Naïve Bayes* (MNB) dan *Multivariate Bernoulli Naïve Bayes* (BNB) untuk mengklasifikasi tweet ke dalam dua kelas, tweet positif dan tweet negatif. Hasil dari penelitian tersebut didapat bahwa akurasi menggunakan metode MNB memiliki akurasi yang sangat tinggi, yaitu 85,33%[5].

Dari penelitian yang dilakukan oleh Faizal (2020) tentang Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma *Support Vector Machine*. Hasil dari penelitian tersebut didapat bahwa dengan *Support Vector Machine*, didapat akurasi sekitar 90%[6].

Berikut merupakan penelitian-penelitian yang sudah dilakukan sebelumnya

Tabel 1 Penelitian yang Sudah Dilakukan Sebelumnya

Judul Penelitian	Peneliti, Tahun	Metode	Hasil Penelitian
Thumbs up? Sentiment Classification Using Machine Learning Techniques	Pang, 2002	Naïve Bayes Classifier (NBC), Maximum Entropy (ME), dan Support Vector Machine (SVM)	Mengklasifikasikan ulasan film ke dalam kelas ulasan positif dan ulasan negatif. Dengan metode NBC didapat akurasi sebesar 80,2%
A novel classification approach based on Naïve Bayes for	Song, 2017	Multinomial Naïve Bayes (MNB) dan Bernoulli Naïve Bayes (BNB)	Mengklasifikasi *tweet* ke dalam kelas *tweet* positif dan *tweet* negatif.

Twitter sentiment analysis			Dengan metode MNB, didapat akurasi sebesar 85,33%
Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma *Support Vector Machine*	Faizal, 2020	*Support Vector Machine* (SVM)	Mengklasifikasikan ulasan aplikasi Ruangguru menjadi ulasan positif dan ulasan negatif. Dengan metode SVM dan dibantu *K-Fold Validation*, didapat akurasi tertinggi sebesar 90,2% untuk nilai k = 6, 9, 10.

Analisis Sentimen

Analisis sentimen atau yang biasa dikenal dengan opinion mining adalah bidang studi yang menganalisis pendapat /opini seseorang, sentimen, perilaku, dan emosi yang diekspresikan secara tekstual [7].

Analisis sentimen dapat digunakan dalam berbagai kemungkinan domain, dari produk konsumen, jasa kesehatan, jasa keuangan, peristiwa sosial dan politik pada pemilu. Kecendrungan penelitian tentang analisis sentimen berfokus pada pendapat yang menyatakan atau menyiratkan suatu sentimen positif atau negatif. Pendapat mewakili hampir semua aktivitas manusia, karena pendapat dapat mempengaruhi terhadap perilaku seseorang. Dalam dunia nyata, bisnis dan organisasi selalu ingin melihat opini publik tentang suatu produk atau jasa [7].

Multinomial Naïve Bayes

Multinomial Naïve Bayes merupakan metode supervised learning yang menggunakan probabilitas. Multinomial Naïve Bayes lebih difokuskan untuk klasifikasi teks[2]. Multinomial Naïve Bayes juga memiliki fitur unik, yaitu hasil yang didapat untuk masing-masing kelas bersifat independen. Ini artinya, dari dokumen satu ke dokumen berikutnya tidak ada keterkaitannya sama sekali sehingga hasil yang didapat murni dari dokumen yang diolah itu sendiri. Perhitungan probabilitas ulasan d yang memiliki kelas c dapat dilihat pada rumus (1)

$$P(c|d) \propto \mathrm{P}(c) \prod_{i=1}^{n_d} P(w_i|c) \quad (1)$$

Keterangan :

P(c|d) = probabilitas suatu kelas c pada dokumen / teks d

P(c) = probabilitas prior c

$P(w_i|c)$ = probabilitas suatu kata pada kelas c

Perhitungan probabilitas sentimen positif atau sentimen negatif dari suatu ulasan yaitu

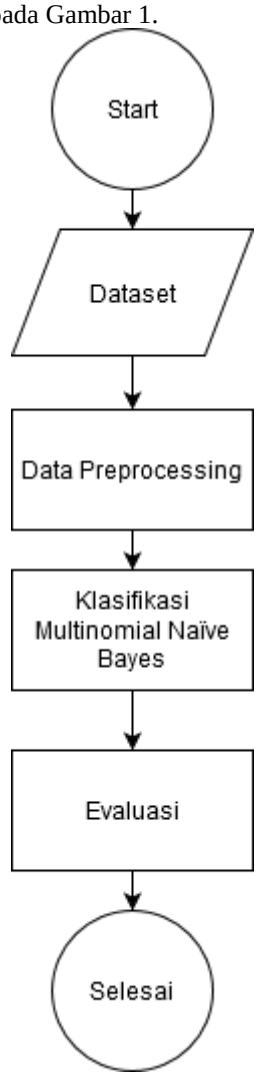
$$P(S|c) = \frac{Banyaknya\ kemunculan\ kata\ sentimen\ positif\ atau\ negatif\ pada\ teks}{Banyaknya\ kemunculan\ kata\ sentimen\ positif\ dan\ negatif\ pada\ teks} \quad (2)$$

Dari rumus (2), dijelaskan bahwa untuk menghitung probabilitas sentimen dari suatu ulasan yaitu dengan cara membandingkan jumlah kemunculan kata sentimen positif atau negatif dengan jumlah seluruh kata sentimen positif dan negatif yang ada pada teks ulasan yang diolah.

3. Sistem yang Dibangun

Gambaran Umum Sistem

Sistem yang dibangun bertujuan untuk mengklasifikasi sentimen dari masukan teks ulasan dengan cara mencari kata yang memiliki sentimen positif atau negatif pada ulasan. Setelah didapat kata sentimen tersebut, sistem akan membandingkan jumlah sentimen kata positif dan kata negatif. Apabila didapat bahwa jumlah sentimen kata positif lebih banyak dibandingkan dengan sentimen kata negatif, maka ulasan tersebut diklasifikasikan sebagai sentimen positif. Namun, jika jumlah kata sentimen nya sama, maka dilihat *rating* bintang yang diberikan. Apabila *rating* nya berupa bintang 5 atau 4, maka ulasan tersebut diklasifikasikan sebagai sentimen positif, dan jika bintang nya 2 atau 1, maka ulasan diklasifikasikan sebagai sentimen negatif. Gambaran umum dari sistem yang dibangun dapat dilihat pada Gambar 1.



Gambar 1 Sistem yang Dibangun

Dataset

Dataset yang digunakan yaitu sebanyak 1500 ulasan 10 aplikasi yang ada pada *Google Play Store* terkecuali ulasan yang memiliki *rating* bintang 3 dan dibagi menjadi 150 ulasan untuk masing-masing aplikasi yang dibagi menjadi 50 ulasan bintang 5, 50 ulasan bintang 1, 25 ulasan bintang 2, dan 25 ulasan bintang 1. Dari 50 ulasan yang diambil dari bintang 5 dan bintang 1, akan diambil 25 ulasan yang akan digunakan pada skenario pengujian kedua yang akan dijelaskan lebih lanjut. Ulasan yang diambil yaitu ulasan yang dianggap bahasa Indonesia oleh *Google Play Store*. 10 aplikasi yang diambil ulasannya yaitu : Lazada, Ruang Guru, Gojek, Grab, Spotify, Tokopedia, Linkaja, Kredivo, LINE Webtoon, dan BukaLapak.

Dataset ulasan yang diambil akan diubah menjadi Bahasa Indonesia untuk menghilangkan beberapa kata campuran Bahasa Inggris yang tercantum, seperti "*good*" akan diubah menjadi "bagus". Serta akan diperbaiki juga beberapa ejaan seperti "baguuusssss" yang diubah menjadi "bagus" dan "trmksh" menjadi "terimakasih". Hal ini dilakukan untuk meminimalisir tingkat *error* yang diakibatkan oleh beberapa kata yang tidak dapat diproses oleh sistem. Untuk *dataset* yang bersifat mentah atau *raw*, akan digunakan untuk pengujian skenario khusus untuk membandingkan performa model klasifikasi yang dibangun.

List kata positif dan kata negatif yang digunakan pada penelitian ini didapat dari penelitian yang sudah dilakukan sebelumnya oleh Wahid pada tahun 2006 dengan judul "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan *Cosine Similarity*".

Preprocessing

Dari *dataset* yang disediakan, akan diolah melalui *preprocessing* terlebih dahulu. *Preprocessing* berguna untuk mengubah bahasa alami menjadi bahasa mesin yang dapat dikenali komputer [9]. Untuk proses ini akan dibantu oleh *library string*, NLTK, dan Sastrawi. Berikut beberapa hal yang dilakukan :

1. *Punctual Removal*
   Pada proses ini dilakukan penghapusan karakter *punctual* atau tanda baca yang dibantu oleh *library string* . Contohnya : .,'"!? dan lainnya.
2. *Lower Case*
   Pada tahap ini, semua huruf pada ulasan diubah menjadi huruf kecil yang dibantu oleh *library string*, contoh *lower case* ditampilkan pada Tabel 2,

Tabel 2 Ilustrasi *Lower Case*

Sebelum Lower Case	Hasil *Lower Case*
Aplikasi nya bagus sekali **S**aya suka **T**erbaik	**a**plikasi nya bagus sekali **s**aya suka **t**erbaik

3. *Stemming*
   Setelah tanda baca dihilangkan, dilakukan *Stemming*, yaitu mengubah semua kata ke versi baku nya. Contoh *Stemming* yaitu mengubah "terbaik" menjadi "baik", "kesalahan" menjadi "salah", dan lain-lain. Proses ini dibantu oleh *library* Sastrawi. Untuk contoh *visual stemming* ditampilkan pada Tabel 3

Tabel 3 Ilustrasi *Stemming*

Sebelum *Stemming*	Hasil *Stemming*
aplikasi nya bagus sekali saya suka **terbaik**	aplikasi nya bagus sekali saya suka **baik**

4. *Tokenization*
   Proses *tokenization* adalah memecah teks menjadi *token* atau potongan lebih kecil. Kalimat ulasan dipecah menjadi kata-kata. Proses ini dibantu oleh *library* NLTK. Contoh pada Tabel 4

Tabel 4 Ilustrasi *Tokenization*

Sebelum *Tokenization*	Hasil *Tokenization*
aplikasi nya bagus sekali saya suka baik	['aplikasi', 'nya', 'bagus', 'sekali', 'saya', 'suka', 'baik']

5. *Stopword Removal*
   Pada proses ini akan dihilangkan setiap kata yang tidak relevan dengan cara proses *stringmatching* dengan *stopword* yang tersimpan pada *library* NLTK. Contoh nya ada pada Tabel 5

Tabel 5 Ilustrasi *Stopword Removal*

Sebelum *Stopword Removal*	Hasil *Stopword Removal*
[‘aplikasi’, ‘nya’, ‘bagus’, ‘sekali’, ‘saya’, ‘suka’, ‘baik’]	[‘aplikasi’, ‘nya’, ‘bagus’, ‘suka’, ‘baik’]

Skenario Pengujian

Dibentuk tiga skenario pengujian berbeda dari *dataset* untuk mengukur performansi dari yang digunakan, yaitu:

- Skenario pertama
  Untuk skenario ini, *dataset* yang digunakan yaitu ulasan yang memiliki *rating* bintang 5 bintang 1. Diambil 50 data ulasan untuk tiap aplikasi yang memiliki *rating* bintang 5 dan 50 data ulasan untuk tiap aplikasi yang memiliki *rating* bintang 1 sehingga untuk tiap aplikasi diolah sebanyak 100 data ulasan. Akurasi untuk masing-masing kelas dipisahkan terlebih dahulu untuk mendapat akurasi *rating* positif dan akurasi *rating* negatif, setelah itu diambil akurasi keseluruhan per aplikasi dan diambil rata-rata untuk seluruh aplikasi yang diolah.
- Skenario kedua
  Untuk skenario ini, *dataset* yang digunakan yaitu ulasan yang memiliki *rating* bintang 5, bintang 4, bintang 2, dan bintang 1. Untuk masing-masing *rating*, diambil 25 data sehingga didapat 100 data ulasan tiap aplikasi. Hasil evaluasi dipisahkan menurut *rating* dan kecocokan kelas yang dihasilkan, setelah itu didapat akurasi per kelas positif dan negatif yang nantinya didapat akurasi keseluruhan dari semua aplikasi yang diolah.
- Skenario ketiga
  Untuk skenario ini, *dataset* yang digunakan yaitu *dataset* ulasan mentah yang tidak dibersihkan secara *manual* seperti mengubah kata Bahasa Inggris menjadi kata Bahasa Indonesia, serta membenarkan ejaan dari ulasan yang diolah. Berbeda dengan dua skenario sebelumnya, *dataset* yang diuji hanya diambil dari salah satu aplikasi saja, yaitu Lazada. Hasil evaluasi yang didapat dari skenario pengujian ini akan dibandingkan dengan hasil evaluasi untuk aplikasi Lazada yang didapat dari dua skenario lainnya.

4. Evaluasi

. Untuk mengevaluasi perfomansi dari klasifikasi yang dilakukan, dibentuk *confusion matrix* yang merupakan metode mengukur perfomansi klasifikasi yang memiliki kelas binomial. Pada *confusion matrix* terdapat istilah actual dan predicted. *Actual* merepresentasikan label dari data asli dan *predicted* merupakan label hasil dari klasifikasi yang dilakukan.

Confusion matrix direpresentasikan pada tabel 6 dibawah :

Tabel 6 *Confusion Matrix*

	Actual Positive	***Actual Negative***
Predicted Positive	*True Positive* (TP)	*False Positive*(FP)
Predicted Negative	*False Negative*(FN)	*True Negative*(TN)

1. *True Positive* (TP) : Nilai yang didapat ketika hasil klasifikasi yang didapat berlabel positif dan data asli nya juga berlabel positif.
2. *True Negative* (TN) : Nilai yang didapat ketika hasil klasifikasi yang didapat berlabel negatif dan data asli nya juga berlabel positif.

3. *False Positive* (FP) : Nilai yang didapat ketika hasil klasifikasi yang didapat berlabel positif sedangkan data asli nya berlabel negatif.
4. *False Negative* (FN) : Nilai yang didapat ketika hasil klasifikasi yang didapat berlabel negatif sedangkan data asli nya berlabel positif.

Recall adalah jumlah ulasan hasil klasifikasi positif yang tepat dibagi dengan jumlah ulasan positif yang ada pada *dataset* [8]. Rumus *recall* dituliskan pada rumus (3)

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

True negative rate (TNR) memiliki definisi yang sama dengan *recall*, hanya saja TNR digunakan untuk menghitung khusus kelas negatif. Rumus TNR terdapat pada rumus (4)

$$True\ negative\ rate = \frac{TN}{TN + FP} \times 100\% \quad (4)$$

Accuracy adalah persentase data hasil klasifikasi yang benar dibandingkan dengan seluruh data yang digunakan. Rumus accuracy terdapat pada rumus (5)

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \quad (5)$$

4.1 Hasil Pengujian

Dari dokumen ulasan yang dimasukkan ke dalam sistem, dihasilkan nilai *recall*, *true negative rate*, dan *accuracy* dari masing-masing skenario. Hasil dari pengujian sistem ditampilkan pada Tabel 7

Tabel 7 Hasil Pengujian

Skenario	*Recall*	*True Negative Rate*	*Accuracy*
Pertama	78,8%	79%	78,9%
Kedua	70,4%	77%	73,7%
Ketiga	68%	88%	78%

Khusus untuk skenario kedua, nilai *recall* yang didapat berasal dari dokumen ulasan bintang 5 dan 4, dan untuk *true negative rate berasal* dari dokumen ulasan bintang 2 dan 1. Untuk hasil *recall* dan *true negative rate* dari masing-masing bintang ditampilkan pada Tabel 8

Tabel 8 Hasil *Recall* dan *True Negative Rate* Detil Skenario Kedua

Bintang	*Recall*	*True Negative Rate*
Lima	76%	-
Empat	64,8%	-
Dua	-	74%
Satu	-	80%

Dan untuk skenario ketiga, nilai *recall* dan nilai *true negative rate* keseluruhan yang dibandingkan dengan masing-masing skenario akan ditampilkan pada Tabel 9 dan untuk nilai *recall* dan nilai *true negative rate* detil berdasarkan pendekatan dataset skenario kedua yang didapat akan ditampilkan pada Tabel 10

Tabel 9 Hasil Pengujian Skenario Ketiga Berdasarkan Pendekatan Skenario Pengujian Lainnya

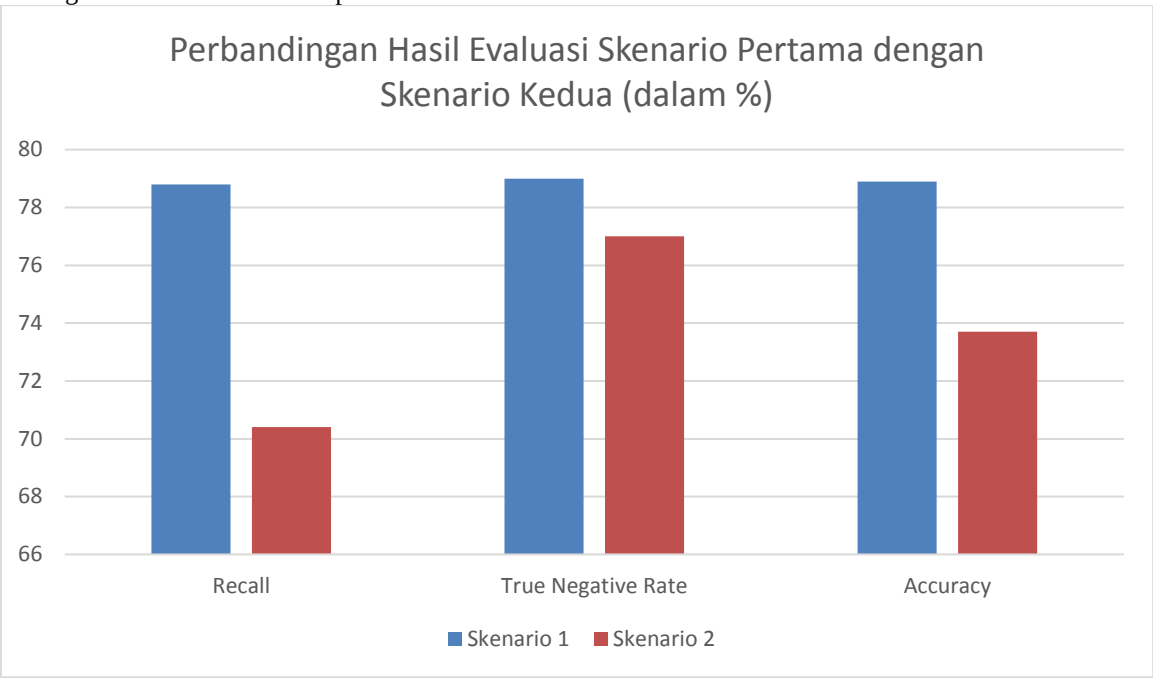
Pendekatan Skenario	*Recall*	*True Negative Rate*	*Accuracy*
Pertama	70%	96%	83%
Kedua	70%	86%	78%

Tabel 10 Hasil *Recall* dan *True Negative Rate* Detil Skenario Ketiga Berdasarkan Pendekatan Skenario Kedua

Bintang	*Recall*	*True Negative Rate*
Lima	76%	-
Empat	64%	-
Dua	-	72%
Satu	-	100%

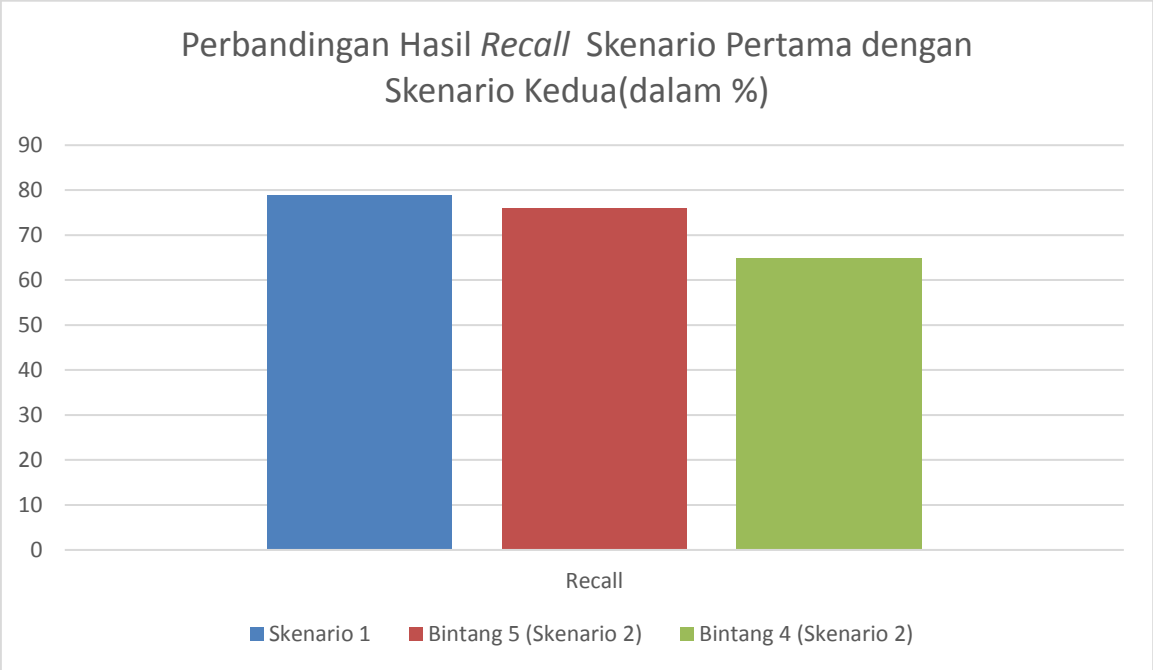
4.2 Analisis Hasil Pengujian

Pada pengujian ini didapatkan hasil *recall*, *true negative rate*, dan *accuracy* dari kedua skenario pengujian yang dilakukan. Pada Gambar 2, ditunjukkan terdapat selisih 8,4% untuk hasil *recall*, selisih 2% untuk hasil *true negative rate*, dan selisih 5,2% untuk *accuracy*. *Accuracy* yang dihasilkan pada skenario kedua turun dibandingkan dengan accuracy dari skenario pertama dikarenakan hasil *recall* dari skenario kedua juga mengalami penurunan dibandingkan recall dari skenario pertama.



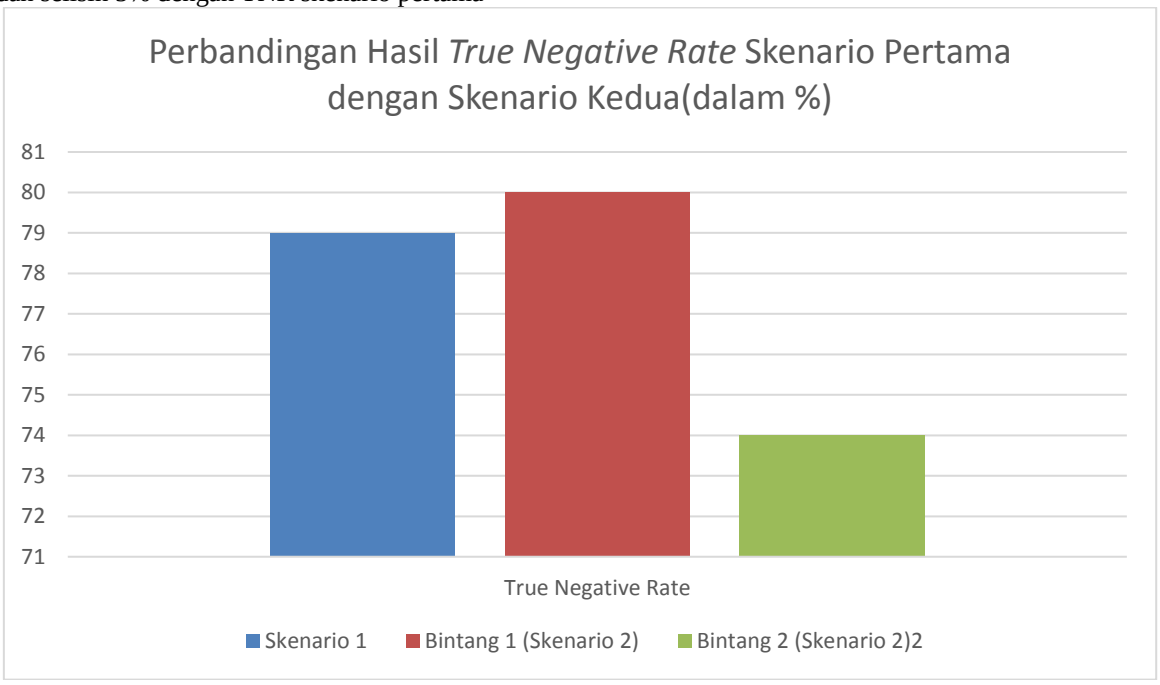
Gambar 2 Hasil Pengujian

Gambar 3 menunjukkan terdapat perbedaan dari hasil *recall* yang dikeluarkan oleh sistem. Hasil *recall* dari skenario pertama memiliki nilai paling tinggi sebesar 78,8% dan menurun perlahan ketika dibandingkan dengan *recall* dokumen ulasan bintang 5 dari skenario kedua sebesar 76% atau selisih 2,8% dan makin menurun ketika dibandingkan dengan recall dokumen ulasan bintang 4 dari skenario kedua sebesar 64,8%, selisih 11,2% dengan *recall* dokumen bintang 5, dan selisih 14% dengan *recall* skenario pertama.



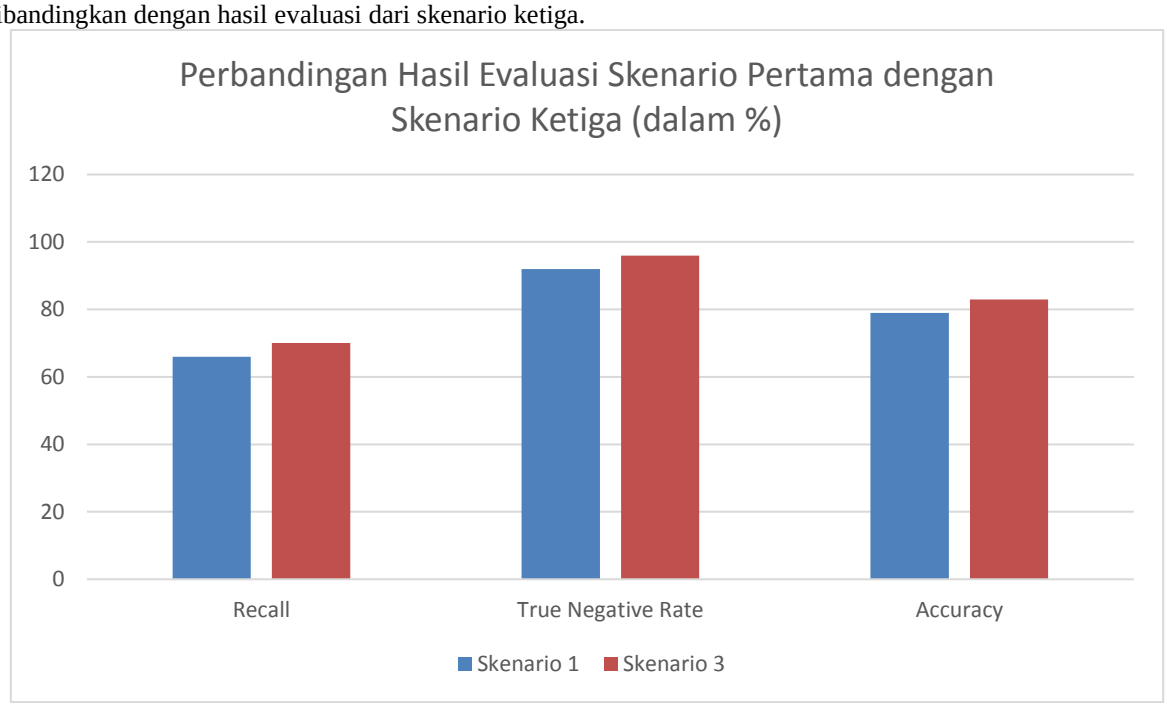
Gambar 3 Hasil Recall

Dan pada Gambar 4, hasil TNR dari sistem juga menunjukkan perbedaan dari skenario pengujian yang dilakukan. hasil TNR dari skenario pertama dihasilkan sebesar 79%. Lalu naik sebesar 1% ketika dibandingkan dengan TNR dokumen ulasan bintang 1 dari skenario kedua sebesar 80%. Dan turun ketika dibandingkan dengan TNR dokumen ulasan bintang 2 dari skenario kedua sebesar 74% selisih 6% dengan TNR dokumen ulasan bintang 1 dan selisih 5% dengan TNR skenario pertama



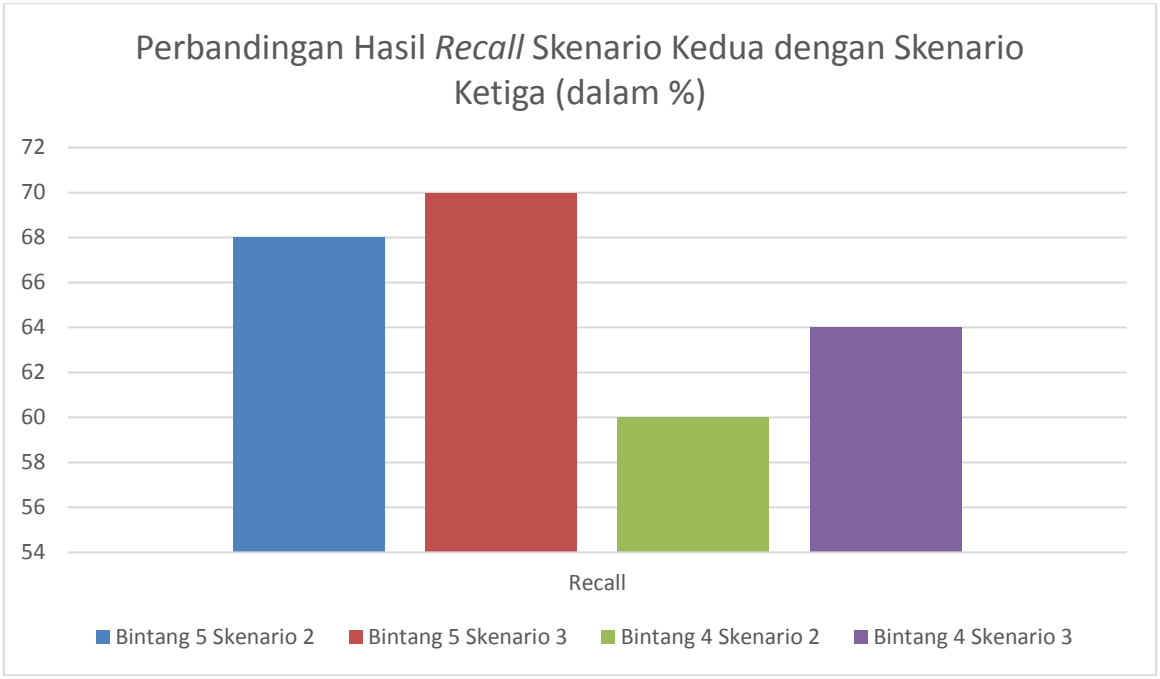
Gambar 4 Hasil *True Negative Rate*

Dari Gambar 5, dilihat nilai *recall* mengalami kenaikan sebesar 4%, nilai *true negative rate* mengalami kenaikan sebesar 4% dan nilai *accuracy* mengalami kenaikan sebesar 4% ketika hasil evaluasi skenario pertama dibandingkan dengan hasil evaluasi dari skenario ketiga.

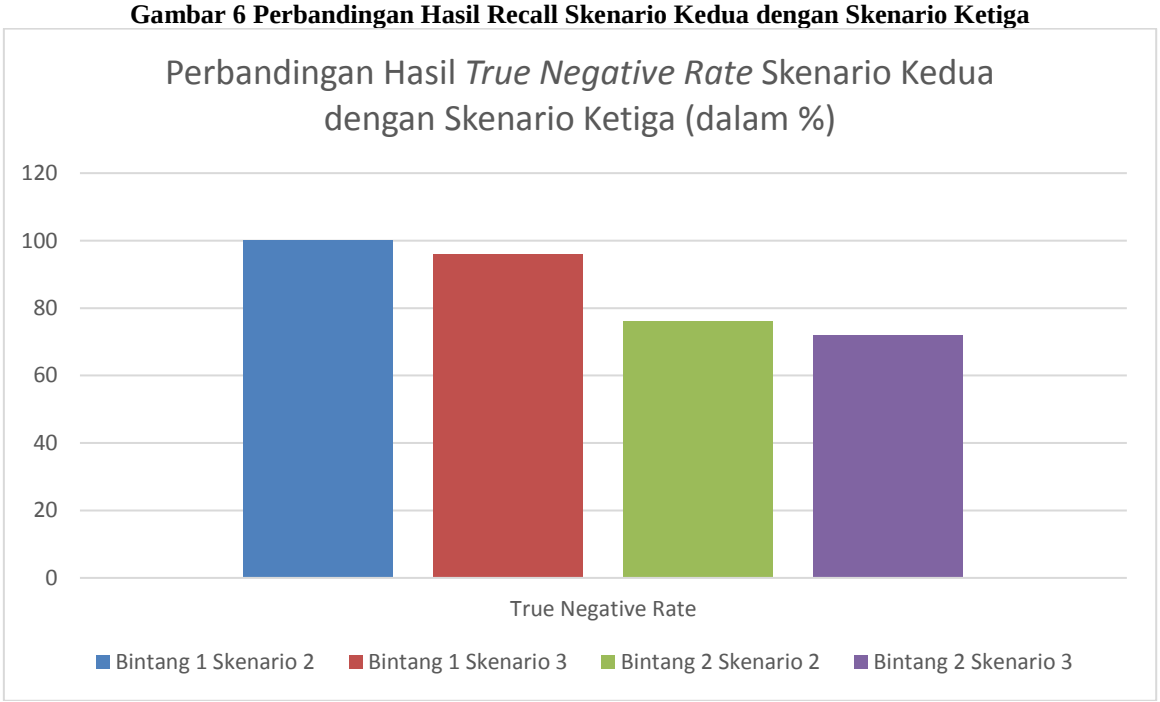


Gambar 5 Hasil Evaluasi Perbandingan Skenario Pertama dengan Skenario Ketiga

Dari Gambar 6, didapat nilai *recall* untuk ulasan bintang 5 mengalami kenaikan sebesar 2% menjadi 70% dan nilai *recall* untuk ulasan bintang 4 mengalami kenaikan sebesar 4% menjadi 64%. Dan dari Gambar 7, didapat juga nilai *true negative rate* untuk ulasan bintang 1 mengalami penurunan sebesar 4% menjadi 96% dan nilai *true negative rate* untuk ulasan bintang 2 mengalami penuruan sebesar 4% menjadi 72%



Gambar 6 Perbandingan Hasil Recall Skenario Kedua dengan Skenario Ketiga



Gambar 7 Perbandingan Hasil True Negative Rate Skenario Kedua dengan Skenario Ketiga

Secara umum, beberapa penyebab hasil *accuracy* yang kurang yaitu karena kurangnya penanganan negasi, contoh : kalimat "tidak bagus" akan diolah oleh sistem menjadi "bagus" karena kata tidak termasuk ke *stopword*. Hal ini dapat menyebabkan perbedaan hasil sentimen yang didapat karena pada ulasan awal, pengguna menyampaikan pengalaman yang buruk, namun oleh sistem dianggap baik karena sistem hanya membaca kata "bagus" nya saja. Dari 1500 ulasan yang diproses, didapat sekitar 48% ulasan yang dipengaruhi oleh hilangnya kata *stopword* "tidak" atau sebanyak 720 ulasan. Dan untuk *stopword* "gak" mempengaruhi sekitar 24% ulasan atau sebanyak 360 ulasan. Penyebab lainnya yaitu beberapa *slang* atau bahasa gaul yang dianggap oleh sistem sebagai *stopword*, atau tidak memiliki nilai sentimen positif atau negatif. Salah satu contoh kata yang dianggap *stopword* oleh sistem yaitu "lanjut" yang dimana apabila dipasangkan dengan "lanjutkan mas!", kata tersebut memiliki sentimen positif, namun karena kata lanjut tidak memiliki nilai sentimen, maka oleh sistem kata tersebut diabaikan. Dari 1500 ulasan yang diproses, didapat sekitar 9,47% ulasan yang dipengaruhi oleh *slang* atau sebanyak 142 ulasan.

5. Kesimpulan

Berdasarkan hasil evaluasi dan analisis yang dilakukan, didapat kesimpulan bahwa :

1. Model klasifikasi sentimen dengan metode MNB dapat menghasilkan akurasi tertinggi 78,9% untuk kasus klasifikasi ulasan berdasarkan teks.
2. Pada halaman *Google Play Store*, didapat setidaknya 78,9% ulasan yang memiliki nilai sentimen yang sesuai dengan *rating* bintang yang diberikan.

Berikut beberapa saran untuk penelitian selanjutnya :

1. Penambahan fitur penanganan negasi untuk memperoleh hasil akurasi yang lebih baik.
2. Pengembangan atau penggunaan *library* selain Sastrawi dan NLTK dapat mempengaruhi hasil *preprocessing* dikarenakan tidak semua kata dapat diolah, terutama Sastrawi untuk proses *stemming* dan NLTK untuk proses *stopword removal*.

Daftar Pustaka

[1] Zendesk; Dimensional Research, "What is the Impact of Customer Service on Lifetime Customer Value?," Zendesk, [Online]. https://www.zendesk.com/resources/customer-service-and-lifetimecustomer-value
[2] Liu, B., Sentiment analysis and Subjectivity. Handbook of natural language processing, 2, pp.627-666, 2010.
[3]B. Pang, L. Lee dan S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," dalam Empirical Methods on Natural Language Processing, New York, 2002
[4] Wahid, D. H., & Azhari, S. N. (2016). Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 10(2), 207-218.
[5] J.Song, K.T. Kim, B. Lee, S. Kim, dan H.Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis" dalam Ksii Transactions On Internet And Information Systems Vol. 11, No. 6, 2017.
[6]F.I. Faizal, T.Mohamad, D.H. Anggit, Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine, Jurnal Bisnis, Manajemen Informatika Vol. 11, No.3, 2020
[7] Feldman, R & Sanger, J. The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, 2007.
[8] D. Kelly et al. Methods for evaluating interactive information retrieval systems with users. Foundations and Trends R in Information Retrieval, 3(1–2):1–224, 2009.
[9] J. M. Zhishuo Liu, Qianhui Shen. Extracting implicit features based on association rules. 2018.
[10] Liu, B. 2012. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies