

Implementasi Algoritma Naïve Bayes Terhadap Klasifikasi Ulasan Aplikasi Tokopedia

Yauw James Fang Dwiputra Harta^{a1}, I Gede Arta Wibawa^{a2}

^aProgram Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Badung, Indonesia

¹jamesfangyauw@gmail.com

²gede.arta@unud.ac.id

Abstract

Tokopedia has an average number of visitors per month which is 149.67 visitors. In addition, the Tokopedia application has been downloaded by more than 100 million users with 6.34 million reviews and received a rating of 4.7 out of 5. This study aims to classify the positive and negative reviews of the Tokopedia application on the Google Playstore using Naive Bayes Algorithm. The results of the 2000 test data testing obtained 842 positive reviews and 1158 negative reviews. This means the percentage for positive reviews is only 42.1%, in contrast to negative reviews of 57.1%. The performance generated in this test against 2000 testing data is 96.19% accuracy value, with a precision value of 1. While in Class Recall the resulting value is 93.45% (positive class: negative). Then for the AUC value is 0.975. It is hoped that these results can help Tokopedia management to improve existing deficiencies so that they can get more positive reviews.

Keywords : Tokopedia, classifier, analisyst sentiment, naive bayes, preprocessing

1. Introduction

Perkembangan *e-commerce* di Indonesia saat ini sangatlah pesat. Jika pada tahun sebelum 2010, kebanyakan masyarakat datang ke toko-toko atau supermarket untuk belanja/membeli barang-barang yang dibutuhkan dan diinginkannya. Tetapi semenjak adanya *e-commerce* ini, masyarakat tidak perlu untuk datang ke toko secara langsung. Cukup dengan memesannya lewat aplikasi *e-commerce* yang dikembangkan oleh perusahaan *e-commerce* tersebut. Bahkan hingga kuartal 1 tahun 2022, menurut hasil statistik yang diterbitkan pada artikel liputan 6, bahwa pemerintah telah mencatatkan statistik jumlah transaksi di *e-commerce* seluruh Indonesia sebesar Rp 108,54 triliun[1].

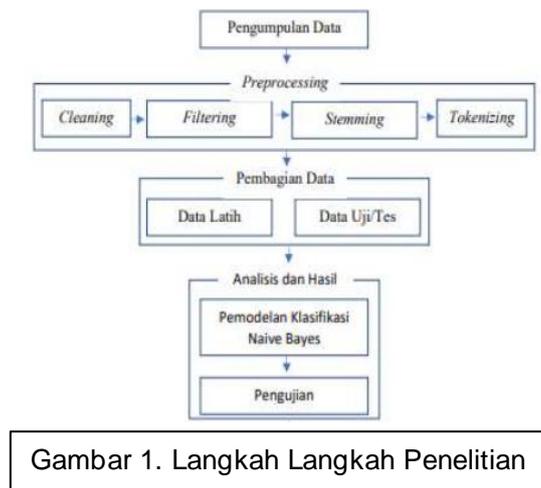
Tentu ini merupakan suatu hal yang positif, karena semakin tinggi angka transaksi pada *e-commerce* tentu ini menandakan bahwa masyarakat Indonesia sudah lebih beradaptasi dengan perkembangan teknologi saat ini khususnya dalam menggunakan aplikasi *e-commerce*. Salah satu *e-commerce* dengan jumlah pengguna terbanyak di Indonesia adalah Tokopedia. Menurut sumber yang dituliskan oleh Vika Azkiya Dihni dikatakan bahwa Tokopedia menjadi yang tertinggi dibandingkan kompetitor lainnya pada tahun 2021 dengan jumlah pengunjung rata-rata tiap bulan adalah 149,67 pengunjung[2]. Selain itu, aplikasi tokopedia sudah didownload oleh lebih dari 100 juta pengguna dengan 6,34 juta ulasan dan mendapatkan rating 4,7 dari 5. Meski memiliki rating yang tinggi di google play store, tetapi tidak semua ulasan dari 6,34 juta ulasan tersebut memiliki arti yang positif, tidak sedikit juga yang bermakna negatif. Ulasan-ulasan negatif tersebut tentu akan bisa menjadi tolak ukur kita mengenai performa dan layanan aplikasi tokopedia, selain tolak ukur juga bisa menjadi pacuan perusahaan tokopedia untuk mengembangkan agar aplikasinya menjadi lebih baik lagi. Maka dari itu lah, penulis merasa perlu untuk melakukan analisis ulasan aplikasi tokopedia dan mengklasifikasikannya kedalam kelompok positif dan negatif.

Pada tahun 2020, hal ini sudah pernah diteliti oleh Afifah Faadilah dengan jurnal yang berjudul "Analisis Sentimen Pada Ulasan Aplikasi Tokopedia Di Google Play Store Dengan Menggunakan Metode Long Short Term Memory"[3]. Penelitian tersebut mengambil ulasan data dengan jumlah 3067 dengan jumlah kelas negatif 2591 dan positif 476. Tentunya data

ulasan tahun 2020 tersebut telah tidak relevan lagi saat ini sehingga penulis mengambil topik ini. Perbedaan penelitian ini dengan penelitian sebelumnya adalah terletak pada metode algoritma yang digunakan, dimana pada penelitian ini menggunakan metode Naive Bayes dengan dataset yang berbeda juga. Hasil akhir dari penelitian ini adalah tingkat prosentase / jumlah ulasan aplikasi tokopedia yang telah diklasifikasikan menjadi kelompok positif dan negatif. Hasil ini akan dapat menjadi manfaat untuk manajemen Tokopedia kedepannya untuk lebih mengembangkan aplikasi tokopedia ini menjadi lebih baik sehingga masyarakat Indonesia tetap memberikan kepercayaannya kepada pihak Tokopedia yang tentu ini menjadi keuntungan juga bagi negara karena tokopedia merupakan perusahaan asli buatan warga Indonesia yang akan bisa memberikan income kepada negara maupun UMKM di *marketplace* Tokopedia.

2. Research Methods

Metode penelitian yang digunakan pada penelitian ini adalah dengan menggunakan metode eksperimen dengan urutan langkah-langkah seperti gambar dibawah ini.



Gambar 1. Langkah Langkah Penelitian

Berikut adalah penjelasan lebih lanjut mengenai diagram alur metode penelitian diatas :

1. Pengumpulan Data

Pengumpulan data yang digunakan dalam penelitian ini dilakukan dengan dua tahap, tahap pertama adalah mengumpulkan data yang akan dijadikan data latih/training dan tahap kedua mengumpulkan data yang akan dijadikan data uji/testing. Data yang diambil adalah ulasan bahasa Indonesia dari aplikasi Tokopedia yang ada di playstore yang dimulai dari 1 Maret 2022 hingga 29 September 2022 dengan total ulasan sebanyak 2000 untuk data testing dan 1000 untuk data latih/training. Data latih/training yang telah dikumpulkan ini diberikan label positif/negatif secara manual. Dalam pengumpulan data ini menggunakan metode web scrapping pada browser chrome dengan menggunakan bantuan extension yaitu dataminer[4].

Contoh :

Redy Erwin
27 September 2022
di android redmi note 7, Kenapa ya aplikasi Tokopedia berat banget, Kadang2 berenti, mau masuk ke apk susah, mau geser cari produk jg gak bisa, aku kira sudah ada perbaikan, ternyata masih belum, tolong dong Tokopedia,, biar bisa aku pertahankan aplikasi nya.

2. Preprocessing

Setelah didapatkan dataset, maka langkah selanjutnya adalah melakukan preprocessing dataset tersebut yang diimplementasikan dengan bahasa pemrograman python. Adapun pada preprocessing ini terdiri dari 4 tahapan yaitu sebagai berikut[5] :

a. Cleaning

Pada proses ini kita melakukan pembersihan pada dataset yang telah kita dapatkan tadi, dimana pembersihan ini meliputi menghapus karakter-karakter selain huruf, menghapus nama/email pengguna, menghapus emoticon, menghapus hashtag (#), menghapus url, dan mengubah semua huruf menjadi huruf kecil.

Contoh :

```
september  
di android redmi note kenapa ya aplikasi tokopedia berat banget kadang berenti  
mau masuk ke apk susah mau geser cari produk jg gak bisa aku kira sudah ada  
perbaikan ternyata masih belum tolong dong tokopedia biar bisa aku  
pertahankan aplikasinya
```

b. Filtering

Pada tahap ini dilakukan penyaringan dengan menerapkan proses untuk mendapatkan kata kata dan melakukan pembuangan terhadap kata kata yang tidak penting. Setelah itu dilakukan juga penghapusan tanda baca yang ada dan juga kata kata yang termasuk *stopwords*. Kriteria dari *stopwords* ini yaitu tidak memiliki makna penting, sering keluar dalam teks . Contohnya seperti kata kata penghubung, waktu, dan lain sebagainya.

Contoh :

```
android redmi note kenapa aplikasi tokopedia berat berenti masuk apk susah  
geser cari produk gak bisa aku kira ada perbaikan ternyata belum tolong  
pertahankan
```

c. Stemming

Setelah didapatkan hasil kata kata dari filtering, maka dilakukanlah stemming. Pada stemming ini kita mengubah kata kata yang memiliki imbuhan menjadi kata kata dasar sesuai EYD.

Contoh :

```
android redmi note kenapa aplikasi tokopedia berat henti masuk apk susah geser  
cari produk tidak bisa aku kira ada baik ternyata belum tolong tahan
```

d. Tokenizing

Setelah didapatkan data berupa kata kata dasar, maka tahap selanjutnya adalah memisahkan kata kata tersebut sesuai dengan spasinya. Proses inilah yang dinamakan sebagai tokenizing.

Contoh :

```
['android', 'redmi', 'note', 'kenapa', 'aplikasi', 'tokopedia', 'berat', 'henti', 'masuk',  
'apk', 'susah', 'geser', 'cari', 'produk', 'tidak', 'bisa', 'aku', 'kira', 'ada', 'baik',  
'ternyata', 'belum', 'tolong', 'tahan']
```

3. Pembagian Data

Mula-mula data diberikan pelabelan secara manual dan setelah itu dilakukanlah pembagian data menjadi 2 yaitu data latih dan data uji. Dimana terdapat 2 kelompok dalam penelitian ini yaitu kelompok data yang pertama adalah jumlah data latih lebih sedikit dari data uji. Sedangkan kelompok data kedua adalah data laih lebih banyak dari data uji.

4. Analisis dan Hasil

Pada tahap ini, peneliti melakukan analisis dengan menerapkan metode naive bayes yang diimplementasikan dengan bahasa pemrograman python. Definisi dari naive bayes yaitu suatu metode untuk mengklasifikasi probalistik yang berdasarkan Teorema Bayes[6]. Rumus dari naive bayes seperti dibawah ini :

$$P(B) = \frac{P(A) \cdot P(A)}{P(B)}$$

Keterangan :

B : Data dengan *class* yang masih belum diketahui

A : Hipotesis data A yang merupakan suatu *class* bersifat spesifik

$P(A|B)$: Probabilitas dari hipotesis A bedasarkan kondisi kasus B (posteriori probability)

$P(A)$: Probabilitas dari hipotesis A (prior probability)

$P(B|A)$: Probabilitas kasus B berdasarkan kondisi hipotesis A

$P(B)$: Probabilitas hipotesis B

Hasil dari ini nanti akan digunakan untuk menentukan ulasan yang termasuk positif dan negatif.

3. Result and Discussion

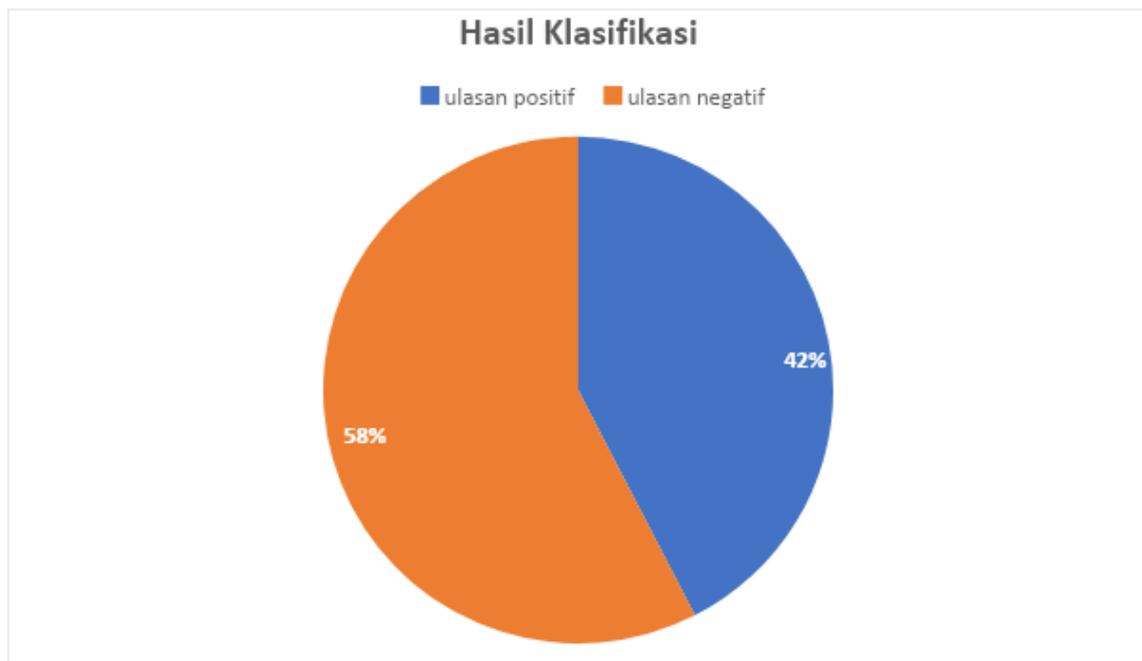
Dataset yang telah dikumpulkan tadi yaitu data latih dan data testing akan dimasukan ke dalam filenya masing masing dengan nama *training_ulasan.xlsx* dan *testing_ulasan.xlsx* yang dimana 2 data set ini akan nilai untuk pembobotan kata, dalam memberikan nilai tersebut dilakukan secara otomatis. Selain pemberian nilai tersebut, juga dilakukan pemberian label yang dilakukan secara otomatis oleh aplikasi pada penelitian ini menggunakan algoritma naive bayes. Untuk data latih/training telah didapatkan label di masing masing datanya karena pada tahapan eksperimen tadi penulis melakukannya secara manual untuk pemasangan label pada data training. Dimana terdapat 500 data yang dilabel positif dan 500 data yang dilabel negatif. Hal ini dilakukan agar data training mempunya kategori atau label sehingga dapat dilakukan proses pelatihan dan pengujian data.

Pada tahapan klasifikasi dengan algoritma berupa naïve bayes, dimana proses ini dimulai dengan melakukan pengukuran kata kata. Pengukuran kata kata dilakukan pada dataset untuk data latih/training. Hasil pada pengukuran ini adalah nilai probabilitas/kemungkinan suatu data positif dan probabilitas suatu data termasuk negatif. Kemudian terkait hasil dari pengukuran pada data training tadi, selanjutnya digunakan dalam menentukan nilai acuan atau patokan ketika mengklasifikasikan ulasan positif dan ulasa/sentimen negatif dari data testing yang menjadi data utama dari penelitian ini. Penentuan sentimen positif atau negatif dilakukan dengan menghitung nilai probabilitas dari data testing, setelah dihitung kemudian nilai tersebut dibandingkan dengan nilai probabilitas yang telah didapatkan dari data training pada proses yang sebelumnya.

Pada pelaksanaan proses ini, metode yang digunakan adalah algoritma naïve bayes yang dapat melakukan klasifikasi ulasan tokopedia secara otomatis oleh aplikasi berbasis bahasa pemrograman python yang dibuat dengan mengimplementasikan algoritma naïve bayes. Pada konsepnya, proses ini dilakukan dengan membandingkan nilai dari data testing

dengan nilai dari data training yang dijadikan patokan untuk klasifikasi. Setelah dilakukan proses tersebut hingga semua kata telah berhasil dibandingkan maka akan didapatkan hasil berupa dataset untuk data testing telah terklasifikasi menjadi ulasan positif dan ulasan negatif. Ulasan yang termasuk klasifikasi positif karena hasil dari algoritma naive bayes pada proses sebelumnya memiliki nilai probabilitas positif yang lebih besar dibandingkan probabilitas negatif. Begitupun sebaliknya untuk ulasan yang termasuk negatif karena hasil dari probabilitas/kemungkinan negatif yang didapatkan pada proses ini lebih besar daripada probabilitas positif.

Hasil dari analisis ulasan-ulasan pada data testing dengan menggunakan aplikasi yang dikembangkan oleh penulis berbasis python yaitu 842 ulasan positif dan 1158 ulasan negatif. Ini berarti persentase untuk ulasan positif hanya 42,1%, berbanding terbalik dengan ulasan negatif sebesar 57,1%.



Data yang telah didapatkan ini perlu dilakukan suatu evaluasi dengan menerapkan suatu pengujian dan untuk mendapatkan hasil dari pengujian dengan menggunakan bantuan *tools* Rapidminer terhadap 2000 data testing ulasan pengguna aplikasi Tokopedia. Pengujian yang dilakukan berdasarkan nilai Class Recal, accuracy, dan terakhir Class Precision pada analisis klasifikasi ulasan pengguna aplikasi tokopedia ini. Dari hasil pengujian ini didapatkan nilai accuracy sebesar 96,19%, kemudian untuk nilai class precision tergolong pada kelas 1. Kemudian terkait class recall didapatkan nilai 93,45%. Kemudian yang terakhir untuk nilai AUC adalah 0,975.

4. Conclusion and Suggestion

Berdasarkan hasil pembahasan diatas, maka terbukti bahwa metode naive bayes dapat digunakan untuk proses klasifikasi ulasan pada aplikasi tokopedia yang terdapat di google play store. Dalam ujicobanya menggunakan data training dan data testing. Data training telah dilabeli secara manual sebelumnya dan data ini yang akan menjadi acuan dari metode untuk mengklasifikasikan data testing termasuk ulasan positif ataupun ulasan negatif. Dalam mengklasifikasikan itu, peneliti menggunakan bantuan aplikasi rapidminer agar ujicoba dapat dilakukan secara *realtime*. Peforma yang dihasilkan pada pengujian diatas terhadap 2000 data testing adalah 96,19% nilai accuracynya dan untuk nilai precision tergolong pada kelas

dengan nilai 1. Kemudian pada kelas class recall didapatkan nilai 93,45%. Kemudian untuk nilai AUC adalah 0,975. Dari 2000 data testing diperoleh 842 ulasan positif dan 1158 ulasan negatif. Ini berarti persentase untuk ulasan positif hanya 42,1%, berbanding terbalik dengan ulasan negatif sebesar 57,1%. Hasil ini tentu berbanding terbalik dengan penilaian pada aplikasi tokopedia yang mendapatkan rating 4,7 dari 5. Tapi ini bisa disebabkan karena data testing yang digunakan juga masih termasuk sedikit. Maka dari itu peneliti berharap kepada peneliti lain bisa menggunakan jumlah data testing yang lebih banyak juga bisa dengan menggunakan metode algoritma yang lain agar mendapatkan nilai performa yang lebih baik dari naive bayes.

References

- [1] "Transaksi E-Commerce Indonesia Rp 108,54 Triliun di Kuartal I-2022," *liputan 6*, 2022. <https://www.liputan6.com/bisnis/read/5032523/transaksi-e-commerce-indonesia-rp-10854-triliun-di-kuartal-i-2022> (accessed Sep. 20, 2022).
- [2] Vika Azkiya Dihni, "Tokopedia, E-Commerce dengan Pengunjung Terbanyak pada 2021," *Databooks*, 2022. <https://databoks.katadata.co.id/datapublish/2022/04/12/tokopedia-e-commerce-dengan-pengunjung-terbanyak-pada-2021> (accessed Sep. 20, 2022).
- [3] A. Faadilah, "ANALISIS SENTIMEN PADA ULASAN APLIKASI TOKOPEDIA DI GOOGLE PLAY STORE MENGGUNAKAN METODE LONG SHORT TERM MEMORY," 2020.
- [4] Gede Putra Aditya Brahmantha and I Wayan Santiyasa, "Sentiment Analysis of the Enforcement of PSBB Part II ," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. Volume 9 No. 2, Sep. 2020.
- [5] "Teknik pre-processing dan classification dalam data science," 2022. <https://mie.binus.ac.id/2022/08/26/teknik-pre-processing-dan-classification-dalam-data-science/> (accessed Sep. 29, 2022).
- [6] Fitri Handayani and Feddy Setio Pribadi, "Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110," *Jurnal Teknik Elektro Vol. 7 No. 1* , 2018.