Optimasi Analisis Sentimen Bea Cukai di Youtube Menggunakan Teknik SMOTE dan Algoritma Naïve Bayes

**Alwin Marcellino1, Yosefa Camilia Moniung2**

1,2Universitas Multi Data Palembang, Jl. Rajawali No. 14, Kota Palembang Sumatera Selatan 30113, telp: (0711) 376 400

3Jurusan Informatika, FIKR UMDP

e-mail: [**1alwinmarcellino@mhs.mdp.ac.id**](mailto:1alwinmarcellino@mhs.mdp.ac.id), [2camiliayosefa@mhs.mdp.ac.id](mailto:2camiliayosefa@mhs.mdp.ac.id)

***Abstrak***

*AbstrakMaksimal 200 kata berbahasa Indonesia dicetak miring dengan Times New Roman 11point. Abstrak harus jelas, deskriptif dan harus memberikan gambaran singkat masalah yang diteliti. Abstrak meliputi* ***alasan pemilihan topik atau pentingnya topik penelitian, metode penelitian dan ringkasan hasil.*** *Abstrak harus diakhiri dengan komentar tentang pentingnya hasil atau kesimpulan singkat.*

***Kata kunci***—3-5 kata kunci, Algoritma A, algoritma B, kompleksitas

***Abstract***

*A maximum 200 word abstract in English in italics with Times New Roman 11 point. Abstract should be clear, descriptive, and should provide a brief overview of the problem studied. Abstract topics include reasons for the selection or the importance of research topics, research methods and a summary of the results. Abstract should end with a comment about the importance of the results or conclusions brief.*

***Keywords***—3-5keywords,AlgorithmA,Balgorithms, complexity

Creative Commons License This is an open-access article under the [CC-BY-CA](https://creativecommons.org/licenses/by-sa/4.0/) license

1. PENDAHULUAN

M

edia sosial adalah wadah masyarakat dalam menyampaikan ide, gagasan, pendapat, dan kritik [9]. Dalam kehidupan sehari-hari, media sosial sangat sering digunakan oleh masyarakat. Menurut We are Social, youtube merupakan peringkat ke-2 website yang sering dikunjungi oleh pengguna internet di Indonesia pada tahun 2024, disusul dengan facebook.com, instagram.com, whatsapp.com, dan lain-lain. Kehadiran Youtube menjadi inovasi baru karena perolehan informasinya melalui bentuk audio dan visual [10]. Karena kelebihannya ini, youtube menjadi wadah aspirasi masyarakat sekaligus faktor cepatnya penyebaran informasi dan berakhir menjadi sesuatu yang *viral*. Salah satu contoh hal yang sedang *viral* belakangan ini adalah permasalahan bea cukai.

Baru-baru ini bea cukai menjadi *viral* karena permasalahan bea masuk yang terlalu tinggi, dimana seorang pria membeli sepatu seharga 10 juta rupiah namun dikenakan bea masuk sebesar 30 juta rupiah (<https://www.detik.com/jatim/berita/d-7296661/viral-tkw-madura-kena-pajak-rp-360-juta-usai-bawa-pulang-emas-3-kg-dari-arab>) serta seorang TKW asal Madura yang terkena bea masuk ratusan juta rupiah saat membawa masuk emas seberat 3kg dari Arab Saudi ke Indonesia (<https://www.cnnindonesia.com/ekonomi/20240423074839-532-1089269/bea-cukai-respons-viral-beli-sepatu-rp10-juta-kena-bea-masuk-rp30-juta>). Kedua permasalahan tersebut menjadi *viral* di berbagai media sosial, salah satunya di Youtube, dan berhasil menuai berbagai pro dan kontra. Maka dari itu, permasalahan ini dipilih untuk dilakukan analisis sentimen dengan sumber data dari Youtube.

Analisis sentimen merupakan suatu teknik atau sistem untuk ekstraksi data teks yang bertujuan untuk mendapatkan informasi mengenai sebuah sentimen apakah positif atau negatif [9]. Analisis sentimen dapat digunakan untuk memperoleh gambaran opini masyarakat terhadap suatu isu yang terjadi. Pada Youtube, opini masyarakat biasanya tertera di kolom komentar dan bisa menghasilkan ratusan hingga ribuan komentar dalam waktu singkat. Hal ini akan memakan waktu yang lama untuk proses klasifikasinya. Maka dari itu, diperlukan algoritma untuk menyelesaikannya. Pada penelitian ini, akan digunakan algoritma *Naïve Bayes Classifier.*

Implementasi analisis sentimen menggunakan Naïve Bayes telah dilakukan oleh [2]. Pada penelitian ini menghasilkan akurasi sebesar 83% menggunakan pembagian data *Hold-Out* dan akurasi sebesar 82% menggunakan metode K-Fold Cross Validation dengan nilai K=10. Selanjutnya, ada pula penelitian yang dilakukan oleh [6] yakni analisis sentimen opini masyarakat terhadap film Ngeri-ngeri Sedap pada Twitter menggunakan Naïve Bayes. Penelitian ini memperoleh akurasi 75% dan membuktikan sentimen masyarakat terhadap film tersebut tergolong netral. Lalu, penelitian dilakukan oleh [8] yang membandingkan metode SVM dan Naïve Bayes dalam analisis sentimen kandidat Gubernur Java Barat 2018-2023 berdasarkan opini masyarakat di Twitter. Penelitian ini menunjukkan bahwa akurasi yang dihasilkan oleh algoritma Naïve Bayes lebih tinggi, yakni 94% untuk pasangan 2DM (Deddy Mizwar and Dedi Mulyadi) dan SVM menghasilkan 75.50% untuk pasangan 2DM juga. Kemudian, menurut penelitian ini, Naïve Bayes sering digunakan dalam klasifikasi teks karena kemudahan dan kecepatannya.

Berdasarkan penjelasan di atas, Naïve Bayes dinilai efektif dalam analisis sentimen terhadap bea cukai. Namun, pada permasalahan ini, akan ada tiga jenis komentar, yaitu positif, negatif, dan netral. Pada kasus ini, mayoritas masyarakat memberi komentar negatif kepada bea cukai. Hal ini menyebabkan ketidakseimbangan proporsi pada dataset. Ketika proporsi dataset tidak seimbang, pada proses training akan menghasilkan model machine learning yang bias pada kelas set yang paling mayoritas [12]. Untuk mengatasi masalah ini, akan digunakan teknik SMOTE (*Synthetic Minority Oversampling Technique).*

SMOTE merupakan metode populer untuk menangani ketidakseimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas [4]. Penerapan SMOTE ini dapat meningkatkan akurasi dalam klasifikasi. Hal ini ditunjukkan oleh penelitian yang dilakukan oleh [4] yakni penerapan SMOTE untuk *imbalance class* menggunakan KNN dalam analisis sentimen dan diperoleh hasil bahwa klasifikasi menggunakan SMOTE menghasilkan akurasi yang lebih tinggi, yaitu 90% dibandingkan hanya menggunakan KNN, yaitu 82%. Ada pula penelitian yang dilakukan oleh [13] yang menghasilkan peningkatan akurasi yang sebelumnya 69% menggunakan metode SVM menjadi 82% setelah dikombinasikan dengan teknik SMOTE.

Berdasarkan penjelasan di atas, akan dilakukan penelitian untuk menerapkan teknik SMOTE dalam analisis sentimen terhadap bea cukai menggunakan metode *Naïve Bayes Classifier*. Penelitian ini diharapkan dapat dijadikan bahan masukan oleh bea cukai maupun masyarakat untuk melakukan evaluasi terhadap permasalahan ini.

2. METODE PENELITIAN

## 2.1 Landasan Teori

Berikut merupakan landasan teori yang digunakan dalam penelitian ini.

## 2. 1.1 Naïve Bayes

Algoritma Naïve Bayes memiliki beberapa metode yang sering digunakan, yaitu Multinimial, Gaussian, dan Bernaulli [9]. Pada penelitian ini, penulis menggunakan metode Multinomial Naïve Bayes. Multinomial Naïve Bayes adalah metode supervised learning yang dirancang khusus untuk mengklasifikasikan data teks. Metode ini memiliki fitur unik, di mana hasil yang diperoleh dari setiap kelas bersifat independent. Hal ini terjadi karena setiap dokumen tidak memiliki keterkaitan satu sama lain, sehingga hasil yang diperoleh murni berdasarkan informasi dari dokemen tersebut saja [2]. Berikut merupakan rumus dari Naïve Bayes.

Keterangan:

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

P(c) = Probabilitas prior c

P(| c) = Probabilitas suatu kata pada kelas c

*2.1.2* *SMOTE*

*Synthetic Minority Oversampling Technique* (SMOTE) adalah metode yang cukup popular yang digunakan untuk menangani ketidak seimbangan kelas pada dataset. Metode ini mnyeimbangkan dataset dengan cara mensintesis sample baru dari kelas minoritas melalui resampling [4].

## 2. 2 Pengumpulan Data

Pengumpulan data (Crawling data) merupakan tahap pertama yang dilakukan dalam penelitian ini. Dataset yang dikumpulkan merupakan komentar-komentar mengenai opini masyarakat terhadap Bea Cukai di aplikasi Youtube. Data ini diperoleh dengan menggunakan situs web netlytic.org. Setelah pengumpulan data, diperoleh dataset sebanyak 1683 data yang akan digunakan untuk proses selanjutnya.

2.3 Labeling Dataset

Pada tahap ini, penulis akan memberikan label secara manual ke dalam dataset yang telah dikumpulkan sebelumnya. Terdapat 3 label yang akan diberikan, yaitu positif, negative, dan netral.

*2.4 Preprocessing*

Terdapat beberapa proses yang dilakukan, yaitu:

*2.4.1 Case Folding*

Pada proses ini, semua huruf pada dataset akan diubah menjadi huruf kecil (lower case) [2].

*2.4.2 Normalisasi*

Proses normalisasi merupakan proses yang menyamakan format data, seperti menghapus karakter khusus dan tanda baca, serta mengubah singkatan atau slang menjadi bentuk yang lebih formal [4].

*2.4.3 Stopword Removal*

Stopword removal merupakan proses untuk menghilangkan kata-kata umum yang tidak memiliki pengaruh yang signifikan dalam analisis sentiment, seperti “dan”, “di”, “ke”, dan kata-kata sejenis lainnya [4].

*2.4.4 Tokenisasi*

Proses tokenisasi merupakan proses untuk memecah teks menjadi potongan-potongan kata yang dipisah berdasarkan karakter spasi [2].

*2.5.5 Stemming*

Stemming merupakan proses untuk mengubah atau mengembalikan kata ke bentuk dasarnya dengan menghilangkan imbuhan yang ada pada suatu kata[2].

*2.5 TF-IDF*

TF-IDF merupakan sebuah metode yang digunakan untuk menghitung bobot atau menilai pentingnya suatu kata dalam sebuah dokumen relatif terhadap keseluruhan korpus[1].

Keterangan:

d: dokumen ke-d

t: kata ke-t dari kata kunci

W: bobot dokumen ke-d terhadap kata ke-t

TF: banyaknya kata yang dicari pada sebuah dokumen

IDF: Inversed Document Frequency

Nilai IDF didapatkan dari:

Dimana:

D: total dokumen

Df: banyak dokumen yang mengandung kata yang dicari

*2.6 Word Cloud*

Word Cloud adalah salah satu alat yang sering digunakan untuk visualisasi data teks. Word cloud menyusun informasi yang relavan dengan menampilkan istilah-istilah kunci. Kata-kata yang ditampilkan dengan ukuran font terbesar menunjukan tingkat kepentingan yang lebih tinggi [11].

*2.7 Pemodelan*

Pada tahap ini, dataset akan dibagi menjadi dua bagian, yaitu data training dan data testing. Penulis menggunakan metode rasio dalam membagi dataset, dimana 85% digunakan untuk data training dan 15% untuk data test. Selanjutnya, proses klasifikasi dilakukan dengan algoritma yang telah ditentukan, yaitu algoritma Naïve Bayes.

*2.8 Evaluasi*

Proses evaluasi dilakukan untuk mengukur performa yang dihasilkan dari proses klasifikasi. Evaluasi akan dihitung dengan menggunakan *confusion matrix.*

## 2. 2.1 Rumus Matematika

Jika anda menggunakan *Word,*gunakan persamaan Microsoft Equation Editor atau*MathType*, ditulis ditengah, dan diberi nomor persamaan mulai dari (1), (2) dst.

 (1)

## 2. 2.3 Pengacuan Pustaka

Pengacuan pustaka dilakukan dengan menuliskan [nomor urut pada daftar pustaka] mis. [1], [1,2], [1,2,3]. Sitasi kepustakaan harus ada dalam Daftar Pustaka dan Daftar Pustaka harus ada sitasinya dalam naskah. Pustaka yang disitasi pertama kali pada naskah [1], harus ada pada daftar pustaka no satu, yg disitasi ke dua, muncul pada daftar pustaka no 2, begitu seterusnya. Daftar pustaka urut kemunculan sitasi, bukan urut nama belakang. Daftar pustaka hanya memuat pustaka yang benar benar disitasi pada naskah.

3. HASIL DAN PEMBAHASAN

Pembahasan terhadap hasil penelitian dan pengujian yang diperoleh disajikan dalam bentuk uraian teoritik, baik secara kualitatif maupun kuantitatif.Hasil percobaan sebaiknya ditampilkan dalam berupa grafik atau pun tabel. Untuk grafik dapat mengikuti format untuk diagram dan gambar.



Gambar 2 Grafik perbandingan 

# Tabel 1 Perbandingan Algoritma A dan Algoritma B

|  |  |  |  |
| --- | --- | --- | --- |
| Algoritma | Waktu Proses | Ketelitian | Memori |
| A | 120 ms | 98 % | 200 KB |
| B | 105 ms | 95 % | 415 KB |

4. KESIMPULAN

Kesimpulan harus mengindikasi secara jelas hasil-hasil yang diperoleh, kelebihan dan kekurangannya, serta kemungkinan pengembangan selanjutnya.

Kesimpulan dapat berupa paragraf, namun sebaiknya berbentuk point-point dengan menggunakan numbering atau bullet.

5. SARAN

Saran-saran untuk untuk penelitian lebih lanjut untuk menutup kekurangan penelitian. Tidak memuat saran-saran diluar untuk penelitian lanjut.

# UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada xxx yang telah memberi dukungan financial terhadap penelitian ini.

DAFTAR PUSTAKA

● **Buku** dengan urutan penulisan: Penulis, tahun, *judul buku* (harus ditulis miring) volume (jika ada), edisi (jika ada), nama penerbit dan kota penerbit .

[1] Castleman, K. R., 2004, *Digital Image Processing*, Vol. 1, Ed.2, Prentice Hall, New Jersey.

● **Buku Terjemahan** dengan urutan penulisan: Penulis asli (nama depan, tengah. (disingkat), belakang. (disingkat)), tahun buku terjemahan, *judul bukuterjemahan* (harus ditulis miring), volume (jika ada), edisi (jika ada), (diterjemahkan oleh : nama penerjemah), nama penerbit terjemahan dan kota penerbit terjemahan.

[2] Gonzales, R., P. 2004, *Digital Image Processing* (*Pemrosesan Citra Digital*), Vol. 1, Ed.2, diterjemahkan oleh Handayani, S., Andri Offset, Yogyakarta.

● **Artikel dalam Buku** dengan urutan penulisan: Penulis artikel, tahun, *judul artikel* (harus ditulis miring)*,* nama editor, *judul buku* (harus ditulis miring)*,* volume (jika ada), edisi (jika ada), nama penerbit dan kota penerbit.

[3] Wyatt, J. C, danSpiegelhalter, D., 1991, *Field Trials of Medical Decision-Aids: PotentialProblems and Solutions*, Clayton, P. (ed.): *Proc. 15th Symposium on ComputerApplications in Medical Care*, Vol 1, Ed. 2, McGraw Hill Inc, New York.

●***Pustaka dalam bentuk artikel dalam majalah ilmiah:***

Urutan penulisan: Penulis, tahun, judul artikel, *nama majalah* (harus ditulis miring sebagai singkatan resminya), nomor, volume dan halaman.

[4] Yusoff, M, Rahman, S.,A., Mutalib, S., and Mohammed, A. , 2006, Diagnosing Application Development for Skin Disease Using Backpropagation Neural Network Technique, *Journal of Information Technology*, vol 18, hal 152-159.

.

● ***Pustaka dalam bentuk artikel dalam seminar ilmiah:***

**Artikel dalam prosiding seminar** dengan urutan penulisan**:** Penulis, tahun, judul artikel, *Judul prosiding Seminar* (harus ditulis miring), kota seminar, tanggal seminar.

[5] Wyatt, J. C, Spiegelhalter, D, 2008, Field Trials of Medical Decision-Aids: PotentialProblems and Solutions, *Proceeding of 15th Symposium on ComputerApplications in Medical Care*, Washington, May 3.

● ***Pustaka dalam bentuk Skripsi/Tesis/Disertasi*** dengan urutan penulisan: Penulis, tahun, judul skripsi, *Skipsi/Tesis/Disertasi* (harus ditulis miring), nama fakultas/ program pasca sarjana, universitas, dan kota.

[6] Prasetya, E., 2006, Case Based Reasoning untuk mengidentifikasi kerusakan bangunan, *Tesis,* Program Pasca Sarjana Ilmu Komputer, Univ*.* Gadjah Mada, Yogyakarta.

●***Pustaka dalam bentuk Laporan Penelitian:***

Urutan penulisan: Peneliti, tahun, judul laporan penelitian, *nama laporan penelitian* (harus

ditulis miring), nama proyek penelitian, nama institusi, dan kota.

[7] Ivan, A.H., 2005, Desain target optimal, *Laporan Penelitian Hibah Bersaing,*Proyek Multitahun, Dikti, Jakarta.

***Pustaka dalam bentuk artikel dalam internet*** *(tidak diperkenankan*

*melakukan sitasi artikel dari internet yang tidak ada nama penulisnya)****:***

● **Artikel majalah ilmiah versi cetakan** dengan urutan penulisan: Penulis, tahun, judul artikel, *nama majalah* (harus ditulis miring sebagai singkatan resminya), nomor, volume dan halaman.

[8] Wallace, V. P., Bamber,J. C. dan Crawford, D. C. 2000. Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminate analysis and artificial neural network. *Journal Physical Medical Biology* , No.45, Vol.3, 2859-2871.

● **Artikel majalah ilmiah versi *online*** dengan urutan penulisan:Penulis, tahun, judul artikel, *nama majalah* ((harus ditulis miring sebagai singkatan resminya), nomor, volume, halaman dan

alamat website.

[9] Xavier Pi-Sunyer, F., Becker, C., Bouchard, R.A., Carleton, G. A., Colditz, W., Dietz, J., Foreyt, R. Garrison, S., Grundy, B. C., 1998, Clinical Guidlines on the identification, evaluation, and treatment of overweight and obesity in adults, *Journal of National Institutes of Health*, No.3, Vol.4, 123-130, :http://journals.lww.com/acsm-msse/Abstract/1998/11001/paper\_treatment\_of\_obesity.pdf.

● **Artikel umum** dengan urutan penulisan: Penulis, tahun, judul artikel, *alamat website* (harus ditulis miring), diakses tanggal …

[10] Borglet, C, 2003,Finding Asscociation Rules with Apriori Algorithm,[http://www.fuzzy.cs.uniagdeburg](http://www.fuzzy.cs.uni-magdeburg)de/~borglet/apriori.pdf, diakses tgl 23 Februari 2007.

**Daftar Pustaka hanya memuat semua pustaka yang diacu pada naskah tulisan, bukan sekedar pustaka yang didaftar. Pustaka ditulis urut kemunculan pengacuan di naskah, bukan urut abjad penulis.**

[1] Castleman, Kenneth R., 2004, *Digital Image Processing*, Vol. 1, Ed.2, Prentice Hall, New Jersey.

[2] Gonzales, R., P. 2004, *Digital Image Processing* (*Pemrosesan Citra Digital*), Vol. 1, Ed.2, diterjemahkan oleh Handayani, S., Andri Offset, Yogyakarta.

[3] Wyatt, J. C, danSpiegelhalter, D., 1991, *Field Trials of Medical Decision-Aids: PotentialProblems and Solutions*, Clayton, P. (ed.): *Proc. 15th Symposium on ComputerApplications in Medical Care*, Vol 1, Ed. 2, McGraw Hill Inc, New York.

[4] Yusoff, M, Rahman, S.,A., Mutalib, S., and Mohammed, A. , 2006, Diagnosing Application Development for Skin Disease Using Backpropagation Neural Network Technique, *Journal of Information Technology*, vol 18, hal 152-159.

[5] Wyatt, J. C, Spiegelhalter, D, 2008, Field Trials of Medical Decision-Aids: PotentialProblems and Solutions, *Proceeding of 15th Symposium on ComputerApplications in Medical Care*, Washington, May 3.

[6] Prasetya, E., 2006, Case Based Reasoning untuk mengidentifikasi kerusakan bangunan, *Tesis,* Program Pasca Sarjana Ilmu Komputer, Univ*.* Gadjah Mada, Yogyakarta.

[7] Ivan, A.H., 2005, Desain target optimal, *Laporan Penelitian Hibah Bersaing,*Proyek Multitahun, Dikti, Jakarta.

[8] Wallace, V. P., Bamber,J. C. dan Crawford, D. C. 2000. Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminate analysis and artificial neural network. *Journal Physical Medical Biology* , No.45, Vol.3, 2859-2871.

[9] Xavier Pi-Sunyer, F., Becker, C., Bouchard, R.A., Carleton, G. A., Colditz, W., Dietz, J., Foreyt, R. Garrison, S., Grundy, B. C., 1998, Clinical Guidlines on the identification, evaluation, and treatment of overweight and obesity in adults, *Journal of National Institutes of Health*, No.3, Vol.4, 123-130, :http://journals.lww.com/acsm-msse/Abstract/1998/11001/paper\_treatment\_of\_obesity.pdf.

[10] Borglet, C, 2003,Finding Asscociation Rules with Apriori Algorithm,[http://www.fuzzy.cs.uniagdeburg](http://www.fuzzy.cs.uni-magdeburg)de/~borglet/apriori.pdf, diakses tgl 23 Februari 2007.