ANALISIS SENTIMEN ULASAN KEBIJAKAN ZONASI SEKOLAH PADA PENERIMAAN SISWA BARU DENGAN TEKS BAHASA INDONESIA MENGGUNAKAN ALGORITMA NAÏVE BAYES

SENTIMENT ANALYSIS OF SCHOOL ZONATION POLICY REVIEW ON ACCEPTANCE OF NEW STUDENTS WITH INDONESIAN TEXT USING NAÏVE BAYES ALGORITHM

Martarheza Marthiyas[1], Budhi Irawan[2], Casi Setianingsih[3]
[123]Prodi S1 Teknik Komputer, Fakultas Teknik Elektro, Universitas Telkom
[1]**martarheza@student.telkomuniversity.ac.id**, [2]**budhiirawan@telkomuniversity.co.id**,
[3]**setiacasie@telkomuniversity.ac.id**

Abstrak

Dunia pendidikan Indonesia setiap tahunnya terus mencoba memperbaiki sistem pendidikannya salah satu yang cukup kontroversial atau ramai di beritakan adalah kebijakan zonasi dalam penerimaan siswa baru yang berlaku pada tahun ajaran baru 2019/2020. Karena kebijakan Zonasi Sekolah, calon murid yang rumah tempat tinggalnya berada di sekitar sekolah dari tingkat SD hingga SMA diprioritaskan untuk berhak menjadi calon murid dibandingkan dengan siswa yang mungkin saja rumah tempat tinggalnya jauh dari sekolah yang di favoritkannya sejak lama, permasalahan kebijakan zonasi sekolah cukup membuat orang tua murid resah, karena mungkin saja tidak bisa bersekolah di sekolah favorit. Oleh karena permasalahan tersebut dirancanglah sistem sentimen analisis ulasan berdasarkan sosial media twitter tentang permasalahan zonasi sekolah dengan menggunakan Metode Naïve bayes. Sistem analisis setimen kebijakan zonasi sekolah yang dirancang akan menghasilkan klasfikasi dari opini-opini masyarakat pengguna twitter dan dapat menyimpulkan kebijakan tersebut merupakan kategori positif, negatif atau netral dari pengguna twitter di Indonesia. Dengan harapan sistem ini dapat menjadi informasi umpan balik dari permasalahan yang sedang dihadapi oleh pendidikan Indonesia. Pada penelitian ini Sistem menghasilkan akurasi sebesar 90.69%, presisi 90.93%, recall 90.69% dan f1 score 90.75% dengan jumlah fitur terbaik yaitu 2000.

Kata kunci : Klasifikasi Teks, Pendidikan, Zonasi Sekolah, Naïve Bayes, sentimen analisis.

Abstract

The world of education in Indonesia continues to try to improve its education system every year, one of which is quite controversial or widely preached is the zoning policy in admitting new students that applies in the new school year 2019/2020. Because of the School Zoning Policy, prospective students whose homes are located around the school from elementary to high school level are prioritized to be eligible to become prospective students compared to students who may have their homes far from the school they have favored for a long time, the problem of school zoning policies is enough to make parents are worried, because they might not be able to go to a favorite school. Because of this problem, a sentiment analysis review system was designed based on social media Twitter on the issue of school zoning using the Naïve Bayes Method. The school zoning policy setiment analysis system designed will produce classifications of the opinions of the Twitter user community and can conclude that the policy is a positive, negative or neutral category of Twitter users in Indonesia. It is hoped that this system can provide feedback on the problems being faced by Indonesian education. In this study the system produces an accuracy of 90.69%, a precision of 90.93%, a recall of 90.69% and an F1 score of 90.75% with the best number of features, 2000.

Keywords: Text Classification, Naïve Bayes

1. Pendahuluan

1.1 Latar belakang

Sosial media adalah salah satunya tempat berpendapat yang cukup banyak digunakan hingga tahun 2019 yang mayoritas berbasis teks untuk mengekspresikan reaksi mereka terhadap topik dari dunia nyata. Contohnya seperti Twitter. aplikasi sosial media Twitter begitu aktif setiap harinya yang dapat dilihat pada *Trending*. Pada bulan Juli lalu dikutip dari media berita *Cable New Network* Indonesia (CNN) 'Sistem Zonasi Sekolah Berujung Protes di Berbagai Daerah' dijelaskan keputusan pemerintah dalam hal ini Kementerian Pendidikan dan Kebudayaan Republik Indonesia (Kemendikbud) akan adanya Zonasi Sekolah menuai protes di berbagai daerah karena dinilai kontroversi ketika nilai siswa yang rendah namun bisa bersekolah di sekolah yang favorit karena kebetulan tempat tinggalnya dekat dengan sekolah tersebut.

Dengan ketersediaan suatu platform seperti twitter, seharusnya pertumbuhan jaringan sosial yang ada sekarang membuat sebuah instansi menggunakan konten dalam media untuk mengetahui ulasan untuk keputusan yang lebih baik[1].

Oleh karena itu dengan adanya tweet seseorang yang menyampaikan ekspresi dari sebuah keresahan harus tetap diulas kembali agar tidak timbul *bias*, karena pengertian sebuah kalimat dapat diartikan banyak apabila yang membaca memiliki sebuah kepentingan, Sistem analisis ulasan yang dirancang dapat dilakukan dalam sebuah kebijakan pelaku bisnis bahkan kebijakan pemerintahan untuk mengetahui dan memprediksi keputusan terbaik.

1.2 Tujuan

Adapun Tujuan dari penelitian tugas akhir ini merancang sistem analisis sentimen ulasan tentang kebijakan sekolah serta mengetahui performansi sistem analisis sentiment ulasan yang dirancang dengan metode klasifikasi Naïve Bayes.

1.3 Rumusan Masalah

Perumusan masalah dari penelitian tugas akhir ini adalah bagaimana mengetahui sistem analisis sentiment ulasan kebijakan zonasi sekolah di Indonesia serta performansi metode Klasifikasi Naïve Bayes.

2. Dasar Teori

Pada bagian ini menjelaskan teori yang dipakai dalam penelitian tugas akhir ini diantara lain adalah sebagai berikut.

2.1 Sosial Media

Pada era yang telah serba digital Media Sosial/Social Media merupakan sebuah tempat untuk berinteraksi antara manusia hingga perkumpulan atau komunitas dan berkomunikasi tentang hal apa pun, karena memiliki sifat yang efektif, transparan, dikatakan efektif karena sesuatu yang di tuangkan dalam sosial media dalam bentuk teks, gambar, maupun video dapat mudah sekali tersebar.

2.2 Sentimen Analisis

Sentimen analisis merupakan studi yang mempelajari perilaku atau emosi sebuah entitas, *event*, atau atribut lainnya[2], sentimen analisis merupakan suatu tugas untuk membagi suatu teks dalam orientasi tertentu seperti kata Positif, kata Negatif dan Netral[3]. Sentiment analisis adalah bidang yang popular pada suatu penelitian karena keuntungannya dalam berbagai macam aspek, mulai dari prediksi penjualan,politik, dan pengambilan keputusan seorang investor[4].

2.3 Text Mining

Text Mining atau yang lebih dikenal dengan *Opinion* Mining suatu cara untuk mengumpulkan sebuah kata-kata dari berbagai platform seperti web, sosial media yang berbentuk teks, dengan berbagai pokok masalah seperti sosial dan politik yang bertujuan membantu mengetahui alasan dan kepentingan seorang pemangku kepentingan saat akan memilih keputusan tertentu[4].

2.4 Pre-prosesing

Pre-Processing merupakan hal yang penting dalam machine learning agar dapat menghilangkan dataset yang dapat mengganggu sebuah akurasi untuk hasil yang maksimal dari sebuah prediksi. Oleh karenanya dilakukan Beberapa yang dilakukan antara lain: case folding, menghapus URL, menghapus Simbol, stemming, menghapus stopword, tokenisasi.

2.4 Fitur Ekstraksi

Fitur ekstraksi diperlukan pada dokumen yang besar untuk memberikan nilai pada kata-kata, salah satunya adalah menggunakan pembobotan Term Weighting TF-IDF, TF sendiri singkatan dari (Term Frequency) dapat diartikan frekuensi kemunculan sebuah term dalam dokumen yang bersangkutan. IDF merupakan singkatan dari Inverse Document Frequency, dengan kata lain kata yang ada pada data akan diberikan bobot dengan menggunakan TF-IDF.

2.4 Seleksi Fitur

Seleksi Fitur merupakan cara mereduksi atau mengurangi fitur-fitur yang tidak relevan pada proses pengklasifikasian. Fitur seleksi Chi square menggunakan teori statistik untuk menguji independen sebuah *term/fitur* dengan kategorinya. Tujuan utama dari penggunaan seleksi fitur adalah menghilangkan fitur yang kurang relevan dalam klasifikasi.

2.4 Klasifikasi Naïve Bayes.

Algoritma Naïve bayes adalah model probabilitas yang memungkinkan kita untuk menghitung ketidakpastian tentang model dengan prinsip menentukan probabilitas, Naïve bayes sering digunakan sebagai klasifikasi, karena kecepatan matematis dan kesederhanaannya. Naïve Bayes *Classifier* memperkirakan akan keberadaan fitur tertentu dikelas tidak terkait dengan keberadaan fitur lainnya[5]. Naïve Bayes merupakan salah satu teknik teks klasifikasi yang umum dan cukup tua pertama kali diinterpretasikan pertama kali tahun 1774 oleh Pierre Simon Laplace, Naïve Bayes begitu sensitif terhadap pemilihan fiturnya, peningkatan penghitungan waktu dan akurasi yang berkurang dapat terjadi karena adanya fitur yang terlalu banyak dalam proses klasifikasinya[6].

3. Perancangan

3.1 Scrapping Data

Dataset yang digunakan diperoleh dari tweet media social Twitter yang berhubungan dengan kata kunci zonasi, zonasi smp dan zonasi sekolah, jumlah dataset yang diperoleh adalah 1287 data yang terbagi atas 3 kelas yaitu 429 Data negatif, 429 Data netral dan 429 Data positif disimpan dalam bentuk Comma Separated Value (CSV)

3.2 Pre-Processing

Dataset yang telah diperoleh kemudian dilakukan pre-processing, dalam pre-processing diolah dalam beberapa tahap diantaranya:

3.2.1 Case Folding

Case folding adalah tahapan mengonversi data string menjadi ke huruf kecil atau lowercase, karena tidak semua data yang diambil mempunyai konsistensi yang sama sehingga harus diubah menjadi huruf kecil agar seragam.

3.2.2 Menghapus URL

penghapusan *Uniform Resource Locator* atau yang biasa dikenal URL berfungsi agar data yang tweet yang diambil bersih dari *hypertext* yang merujuk ke sebuah web. Seperti (www.url.com), (url.situs@situs.com) karena teks tersebut akan dianggap sebuah kata yang mengganggu dan tidak mengartikan sesuatu dalam teks sentimen.

3.2.3 Menghapus Angka

Proses menghapus angka bertujuan untuk menghapus data yang kurang memiliki makna dan kurang relevan karena jika dibiarkan akan menjadi noise yang akan menambah kosa kata yang tidak penting pada dataset.

3.2.4 Menghapus special karakter

Proses menghapus special karakter bertujuan menghapus beberapa simbol *smiley* dan simbol unik yang umum dipakai pada sosial media umumnya simbol ini terbentuk dari karakter non alphabet.

3.2.5 Menghapus Tanda baca

Proses penghapusan tanda baca bertujuan agar mesin dapat membaca data latih tanpa adanya tanda baca yang tidak bermakna seperti (,),(!) dan lainnya.

3.2.6 Stemming

Proses bertujuan mengganti kata-kata dalam tweet yang tidak baku dengan kata baku atau bentuk awalnya yang ada dalam Bahasa Indonesia. Tujuan nya agar mengurangi ambiguitas yang dapat tertangkap oleh sistem.

3.2.7 Menghapus stopword

Tahapan ini dilakukan untuk menghapus kata henti yang kurang ada arti pada sebuah kalimat atau tidak ada artinya.

3.2.8 Melakukan Tokenisasi

Tahapan ini merupakan pemotongan panjang string input yang berupa kalimat menjadi sebuat penggalan kata, hal ini dilakukan agar pada saat diklasfikasi mesin mempelajari kata bukan kalimat.

3.3 Pembobotan Kata

Tahap pembobotan Text Document yang dimaksud merupakan hasil dari text yang telah dilakukan *pre-processing* dan menjadi input untuk menghitung bobot dari setiap kalimat yang ada pada data. Bobot dari setiap data yang dihasilkan merupakan hasil dari perkalian antara TF dan IDF, TF sendiri dapat diartikan jumlah term pada data, sedangkan IDF adalah jumlah term pada seluruh dokumen yang ada pada data. Program akan menghitung nilai dari setiap kata.

$$w_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

Dimana :

$$idf_t = \log\left(\frac{N}{df_t}\right) + 1 \quad (2)$$

dan TF sendiri merupakan jumlah kata/term yang dicari dalam dokumen

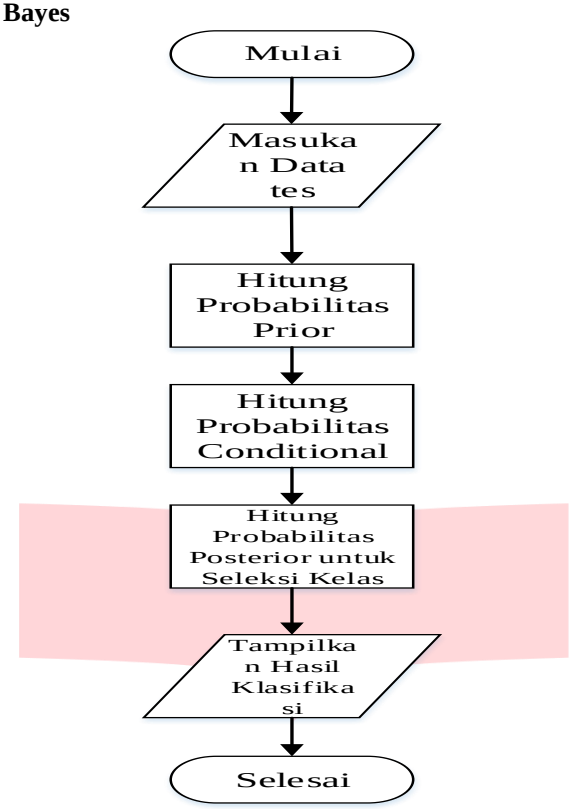
3.4 Seleksi Fitur Chi Square

data yang sudah diperoleh pada text pre-processing dan diberikan nilai dengan TF-IDF, selanjutnya akan disesuaikan kelas dan atribut yang akan membentuk tabel kontingensi. Tabel Kontingensi berfungsi penentuan nilai Chi Square. Setelah tabel kontingensi terbentuk, nilai chi square akan terlihat dengan menggunakan formula

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (3)$$

Kemudian nilai fitur chi square diurutkan terbesar hingga terkecil dan dibatasi dengan jumlah fitur yang ingin digunakan. Apabila fitur tidak termasuk dalam kriteria jumlah fitur yang digunakan maka fitur dibuang.

3.5 Klasifikasi Naïve Bayes



Gambar 3.1 Naïve Bayes

Pada gambar 3.1 digambarkan, setelah fitur diseleksi, selanjutnya data akan diklasfikasi dengan algoritma Naïve bayes, dengan menghitung probabilitas Prior/Kelas, kemudian menghitung Probabilitas kondisional, dan hitung probabilitas posterior untuk menentukan kelas dari data uji. Dengan rumus seperti berikut:

Probabilitas Prior

$$P(c) = \frac{Nc}{N} \quad (4)$$

Probabilitas Conditional

$$P(w|c) = \frac{Hitung\ (w,c)+a}{Hitung(c)+a\ |v|} \quad (5)$$

CMap(Class maximum posterior)

$$C_{Map=\arg\max [\log P(c)+\log P(w1,w2,w3\ldots wn|c)]} \quad (6)$$

4. Pengujian dan implementasi

4.1 Pengujian Partisi Data

Pengujian ini memiliki fungsi untuk mengetahui kinerja algoritma Naïve Bayes dalam mengklasifikasikan data test sesuai kelas yang sudah ditentukan pada dataset, pada dataset akan dibagi dengan beberapa partisi yang salah satu bagian partisinya akan menjadi data testing dan sisanya menjadi data training, partisi dilakukan sebagai bentuk untuk mengukur kesamaan hasil keputusan dengan data yang sudah diberi label.

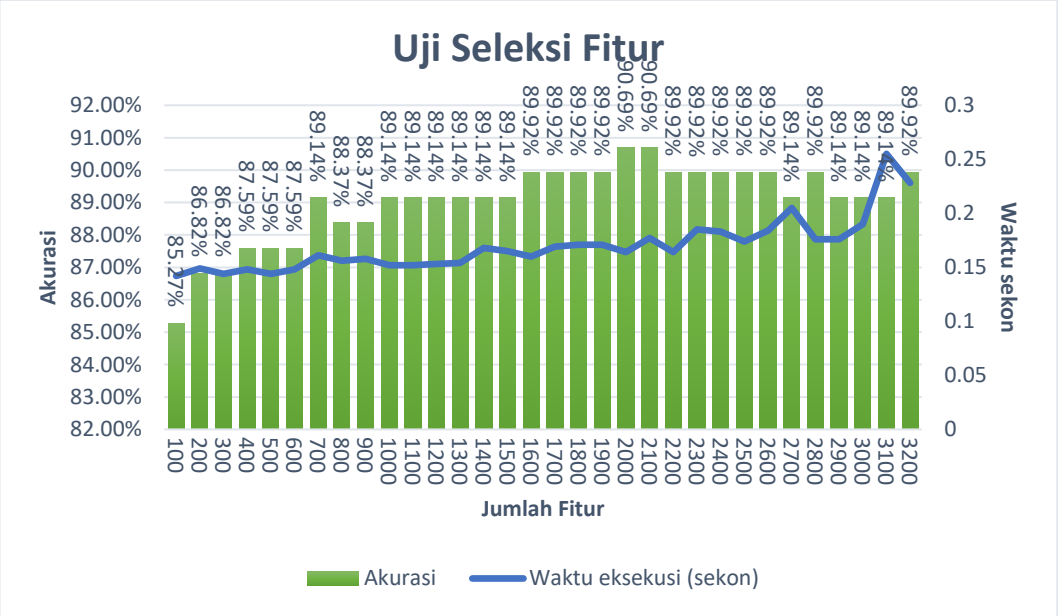
Tabel 4.1 Hasil Uji Partisi Data

Pengujian	Data Latih	Data Uji	Akurasi(%)	Presisi(%)	Recall(%)	F1 score(%)
1	50%	50%	77.63	77.93	77.63	77.71
2	60%	40%	78.83	79.59	78.83	78.92
3	70%	30%	82.17	82.74	82.16	82.17
4	80%	20%	86.82	87.32	86.81	86.90
5	90%	10%	89.92	90.21	89.91	89.99

Pada Tabel 4.1 diatas Penggunaan partisi 90 : 10 menghasilkan Partisi terbaik nilai akurasi = 89.92%, presisi= 90.21%, recall = 89.91%, f1 score = 89.99%, dibandingkan penggunaan partisi 50 : 50 hal tersebut terjadi karena semakin banyak variasi data yang di latih maka semakin banyak juga kata yang dikenal oleh mesin sehingga dapat meningkatkan akurasi untuk memprediksi data. hasil dari partisi terbaik akan terus dipakai ke pengujian selanjutnya.

4.2 Pengujian Seleksi Fitur

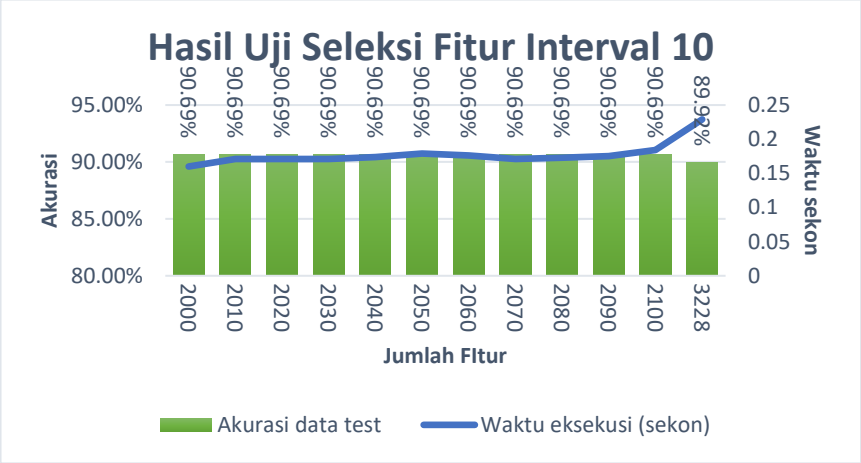
Pada pengujian fitur seleksi ini, akan digunakan jumlah fitur yang berbeda-beda pada tiap percobaan untuk mengetahui perbedaan akurasi dan respon waktu dari setiap jumlah fitur yang diuji, data yang digunakan adalah data partisi terbaik dari pengujian sebelumnya yaitu 90% data training dan 10% data testing.



Gambar 4.1 Uji seleksi fitur

Dalam gambar grafik 4.1 terlihat fitur yang berjumlah 3228 dengan akurasi 89.92% dapat meningkat akurasinya dengan jumlah fitur 2000, akurasi dengan jumlah fitur 2000 mengalami peningkatan menjadi 90.69%, hal ini disebabkan karena sistem banyak mempelajari sebuah kata yang sama dengan kelas sentimen berbeda, fitur tersebut dianggap kurang relevan dan dapat mempengaruhi akurasi[6], sedangkan pemakaian fitur yang terlalu sedikit menyebabkan kurangnya kosa kata dalam data training sehingga dapat terjadi misklasifikasi dan menurunnya akurasi dalam memprediksi data uji sedangkan jika mesin terlalu banyak mempelajari banyak fitur, mesin akan mempelajari banyak fitur yang kurang relevan sehingga menurunkan akurasi prediksi. Pada waktu eksekusi setiap penambahan 100 fitur tidak mengalami perbedaan waktu yang signifikan yaitu sekitar 0.01 sekon namun jika dibandingkan jumlah fitur terkecil pada pengujian ini, jumlah fitur

100 dan jumlah fitur 3231 waktu eksekusi cukup mengalami perbedaan waktu yaitu hingga 0.1 sekon sehingga terbukti penggunaan jumlah fitur yang lebih banyak dapat memperpanjang waktu eksekusi. Lalu diuji kembali 2000 hingga 2100 dengan interval 10.

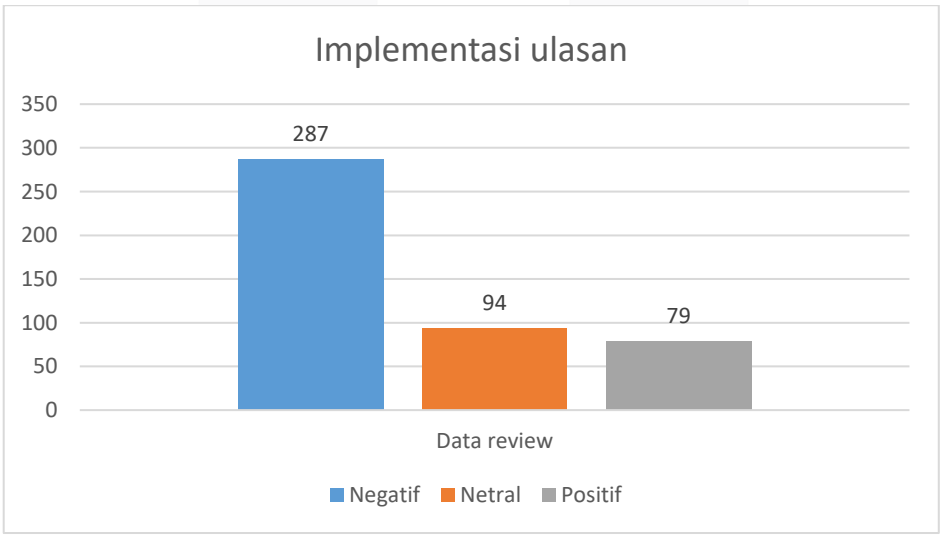


Gambar 4.2 Uji seleksi fitur interval 10

Dari gambar 4.2 diatas dapat disimpulkan jumlah fitur paling optimal untuk dataset adalah 2000 dengan akurasi 90.69% dan waktu eksekusi sistem 0.160 sekon.

4.5 Implementasi Data Uji

Untuk mengimplemtasikannya diambil data twitter terbaru yang berkaitan dengan zonasi sekolah mulai dari 9 Juni 2020 hingga 16 Juni 2020 dan didapatkan jumlah data sebanyak 460 data kemudian dilakukan klasifikasi.



Gambar 4.3 Implementasi ulasan

Pada gambar 4.3 diatas dapat disimpulkan pengguna twitter terhadap kebijakan zonasi sekolah dengan persentase 62.39% Negatif, 20.43% Netral dan 17.17% Positif, mayoritas adalah negatif atau tidak menyukai kebijakan tersebut.

5. Kesimpulan

Berdasarkan pengujian yang dilakukan, diperoleh akurasi terbaik pada partisi 90% data latih dan 10% data uji, fitur yang digunakan sebanyak 2000 dengan akurasi sebesar 90.69%, presisi 90.93%, recall 90.69% dan f1 score 90.75% dan waktu peatihan 0.160 sekon.

Daftar Pustaka:

[1] L. Dini, U. Sekolah, T. M. Informatika, D. Komputer, N. Mandiri, and R. S. Wahono, "Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes," *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.

[2] D. A. Kristiyanti, A. H. Umam, M. Wahyudi, R. Amin, and L. Marlinda, "Comparison of SVM Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based on Public Opinion on Twitter," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–6, 2019.

[3] Dinda Ayu Muthia, "Analisis Sentimen Pada Review Buku Menggunakan Algoritma Naive Bayes," vol. XVI, no. 1, pp. 8–16, 2014.

[4] W. Zhang and F. Gao, "An improvement to naive bayes for text classification," in *Procedia Engineering*, 2011.

[5] A. Goel, J. Gautam, and S. Kumar, "Real time sentiment analysis of tweets using Naive Bayes," *Proc. 2016 2nd Int. Conf. Next Gener. Comput. Technol. NGCT 2016*, no. October, pp. 257–261, 2017.

[6] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Syst.*, 2012.