3rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)
30 Agustus 2023 – Jakarta, Indonesia
Volume 2, Nomor 2, September 2023 - ISSN 2962-8628 (*online*)

# PENERAPAN ALGORITME *K-NEAREST NEIGHBORS* (KNN) UNTUK MENGANALISIS SENTIMEN MASYARAKAT TERHADAP KEBIJAKAN MASUK SEKOLAH PUKUL 5 PAGI

**Aina Fatihah[1*], Haris Munandar[2]**

[1,2]Fakultas Teknologi Informasi, Teknik Informatika, Universitas Budi Luhur, Jakarta, Indonesia

Email: [1*]1911502019@student.budiluhur.ac.id, [2]haris.munandar@budiluhur.ac.id
(* : corresponding author)

**Abstrak-**Penelitian ini membahas kebijakan masuk sekolah pukul 5 pagi di Kupang, NTT, yang menjadi perhatian publik karena kontroversi yang muncul setelah diumumkan oleh Gubernur NTT, Bapak Victor. Video pembahasan kebijakan tersebut menampilkan Gubernur NTT telah menyebar luas di media sosial, khususnya di platform Youtube, dan mengundang berbagai respons dari masyarakat yang pro dan kontra terhadap kebijakan tersebut. Kebijakan masuk sekolah pukul 5 pagi telah menarik perhatian dan mendapat banyak kritik. Namun, masalah pendidikan di wilayah ini tidak dapat diatasi hanya dengan mengubah waktu masuk sekolah. Permasalahan pendidikan yang ada, seperti keterbatasan sarana dan prasarana, serta akses keterbatasan terbatas terhadap pendidikan berkualitas, kualitas terbilang cukup rendah, perlu ditanggapi secara serius. Adapun manfaat dari penelitian ini untuk mengklasifikasikan sentimen masyarakat berdasarkan komentar youtube pada *channel* CNN Indonesia yang membahas tentang kebijakan masuk sekolah pukul 5 pagi, sehingga diperoleh gambaran sentimen masyarakat terkait pandangan masyarakat terhadap kebijakan tersebut. Penelitian ini menggunakan algoritme *K-Nearest Neighbors* (KNN) untuk memahami sikap dan persepsi masyarakat terhadap kebijakan masuk sekolah pukul 5 pagi. Data yang digunakan berasal dari komentar di *channel* youtube CNN Indonesia terkait dengan topik penelitian ini. Hasil penelitian menunjukkan bahwa sebagian besar opini publik menentang kebijakan ini. Hasil pengujian dengan menggunakan *confusion matrix* dan nilai k=11, diperoleh akurasi sebesar 73.68%, presisi sebesar 75%, *recall* sebesar 17.65%, dan *f1-score* sebesar 28.58%.

**Kata Kunci:** *K-Nearest Neighbors*, Analisis Sentimen, *Text Mining, Youtube.*

# *APPLICATION OF K-NEAREST NEIGHBORS ALGORITHM TO ANALYZE PUBLIC SENTIMENT ON SCHOOL ADMISSION POLICY AT 5 AM*

**Abstract-** This study discusses the policy of entering the school at 5 am in Kupang, NTT, which became public attention because of the controversy that arose after it was announced by the Governor of NTT, Mr. Victor. The video of the policy discussion featuring the Governor of NTT has spread widely on social media, especially on the Youtube platform, and invited various responses from the public who are pro and con to the policy. The 5 a.m. school entry policy has attracted attention and received a lot of criticism. However, the education problems in this region cannot be solved simply by changing the time of school entry. Existing educational problems, such as limited facilities and infrastructure, as well as limited access to quality education, of fairly low quality, need to be taken seriously. The benefit of this study is to classify community sentiment based on YouTube comments on the CNN Indonesia channel that discuss the school entry policy at 5 am, so that a picture of community sentiment related to public views on the policy is obtained.This study used the K-Nearest Neighbors (KNN) algorithm to understand people's attitudes and views toward the 5 a.m. school entry policy. The data used came from comments on CNN Indonesia's YouTube channel that were relevant to this research topic. The results showed that the majority of public sentiment towards this policy was negative. From the test results using confusion matrix and k = 11 value, accuracy of 73.68%, precision of 75%, recall of 17.65%, and f1-score of 28.58%.

***Keywords***: K-Nearest Neighbors, *Sentiment Analysis*, *Text Mining*, *Youtube.*

## 1. PENDAHULUAN

Dalam rangka meningkatkan kualitas pendidikan untuk merespon era digitalisasi, pembangunan infrasturktur merupakan hal yang penting dan bahkan menjadi keharusan terutama insfrastuktur yang berbasis digital di zaman milenial ini [1]. Seperti yang terjadi akhir-akhir ini sekolah di Kupang, NTT, menjadi sorotan publik dikarenakan Gubernur NTT Bapak Viktor mengumumkan pemberlakuan adanya penerapan aturan kebijakan masuk sekolah pada pukul 5 pagi. Dengan beredarnya video gubernur NTT yang membahas tentang kebijakan masuk sekolah pukul 5 pagi, banyak masyarakat menyampaikan pendapat dan kritiknya melalui sosial youtube. Maka dari itu

dengan banyaknya komentar pada video tersebut menimbulkan permasalahan pro dan kontra terhadap kebijakan tersebut. Namun, masalah pendidikan di Nusa Tenggara Timur, Kupang, tidak dapat diselesaikan melalui perubahan jam masuk sekolah. Ada tiga masalah pendidikan di NTT. Sarana prasarana pendidikan terbatas, akses pendidikan yang terbatas, kualitas terbilang cukup rendah. Adapun manfaat dari penelitian ini untuk mengklasifikasikan sentimen masyarakat berdasarkan komentar youtube pada *channel* CNN Indonesia yang membahas tentang kebijakan masuk sekolah pukul 5 pagi, sehingga diperoleh gambaran sentimen masyarakat terkait pandangan masyarakat terhadap kebijakan tersebut. Kebijakan ini dijadwalkan mulai berlaku tahun depan, namun beberapa sekolah yang sudah menerapkan kebijakan tersebut, meski aturan tersebut sudah diterapkan tetapi tidak semua sekolah di NTT sudah menjalankannya. Oleh karena itu, penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap kebijakan tersebut dengan menggunakan algoritme *K-Nearest Neighbors* (KNN). Selain itu, penelitian ini juga merancang model untuk menganalisis sentimen terhadap kebijakan tersebut berdasarkan opini komentar dari youtube pada *channel* yang diteliti, dan mengimplementasikan rancangan yang dibuat ke dalam aplikasi berbasis *website*.

Menggunakan *Text Mining* untuk menganalisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen. Metode ini dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk dilakukan analisis keterhubungan antar dokumen [2].

Dalam penelitian ini mengggunakan *K-Nearest Neighbors*, Algoritme *K-Nearest Neighbors* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi mempresentasikan fitur dari data [3]. Hasil dari klasifikasi KNN kemudian dihitung nilai akurasi menggunakan *Confusion matrix* adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi. Dalam pengujian keakuratan hasil percarian akan di evaluasi nilai *recall*, *precision*, *accuracy*, dan *error rate* [4].

a. Akurasi

Akurasi adalah kesesuaian nilai hasil prediksi pengujian dengan nilai aktual (*ground thruth*) yang dibandingkan [5].Rumus menghitung akurasi akan dijelaskan pada persamaan (1).

$$\frac{TP+\ TN}{TP+FN+FP+TN} * 100 \qquad (1)$$

b. *Precision*

*Precision* adalah tingkat ketepatan antara informasi yang diminta oleh *user* dengan hasil jawaban yang diberikan oleh sistem. Dengan kata lain adalah perhitungan untuk menolak dokumen yang tidak relevan dalam dokumen [5]. Rumus untuk menghitung *precision* akan dijelaskan pada Persamaan (2).

$$\frac{TP}{TP+FP} \qquad (2)$$

c. *Recall*

*Recall* adalah jumlah kesesuaian informasi yang didapatkan dari hasil percobaan berdasarkan sudut pandang label atau kelas yang digunakan. Singkatnya pada *recall* adalah perhitungan untuk menemukan semua dokumen yang relevan[5]. Rumus untuk menghitung *recall* akan dijelaskan pada Persamaan (3).

$$\frac{TP}{TP+FN} \qquad (3)$$

d. *F-Meansure*

*F-Measure* adalah bobot *harmonic mean* pada *recall* dan precision. Jadi perhitungan antara *Precision* dan *Recall* [5]. Rumus untuk menghitung *f-measure* akan dijelaskan pada Persamaan (4).

$$\frac{2*P\ x\ R}{P+R} \qquad (4)$$

Penelitian ini dalam pemodelan datanya menggunakan fitur *Countvectorizer* untuk mengekstrak kalimat-kalimat dalam dokumen ke dalam satu kata yang menyusunnya, dan menghitung seberapa sering setiap kata yang hadir dalam setiap dokumen. Setiap dokumen diwakili oleh *vector* yang ukurannya sama dengan jumlah kata[6].

Dalam mendapatkan data sesuai dengan topik yang dibawakan dengan memakai media sosial *youtube*. Setiap pengguna youtube dapat membuat channel sendiri yang memungkinkan penggunanya untuk mengunggah video. Pengguna harus membuat konten video yang menarik agar pengguna lain mempunyai rasa ingin tahu dan menonton video yang di-upload. Ketika menonton konten video di youtube, pengguna merespon video tersebut dengan memberikan bermacam-macam komentar [7]. Analisis sentimen komentar pengguna youtube dalam penelitian ini bertujuan untuk memudahkan para kreator youtube mengetahui jenis video yang diminati pengguna youtube, tanpa harus membaca satu persatu komentar yang diberikan [8].

Adapun tujuan dari pembuatan sistem text mining ini untuk menganalisis sentimen masyarakat. Sentimen analisis merupakan studi yang mempelajari perilaku atau emosi sebuah entitas, event, atau atribut lainnya [9], sentimen analisis merupakan suatu tugas untuk membagi suatu teks dalam orientasi tertentu seperti kata Positif,

kata Negatif dan Netral [10]. Adapun beberapa penelitian yang sudah dilakukan dalam melakukan klasifikasi sentimen terhadap beberapa penelitian sebelumnya, analisis sentimen yang dilakukan menggunakan media sosial facebook yang membahas kepribadian pengguna pada media sosial. Penelitian ini menggunakan algoritme *K-Nearest Neighbor*. Dengan metode tersebut dilakukan pengklasifikasian untuk mengetahui golongan mana yang akan masuk ke dalam kepribadian *Big Five Personality* yaitu *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* dan *Neuroticism* yang dilakukan oleh Putra & Wardani pada tahun 2020. Pada penelitian analisis sentimen untuk mengklasifikasikan tanaman herbal dan perbandingan hasil kinerja menggunakan metode *K-Nearest Neighbor* dan *Local Binary Pattern Histogram*, untuk menghitung akurasi menggunakan *confusion matrix* yang dilakukan oleh Isman, Ahmad & Latief pada tahun 2021. Pada penelitian ini mempresentasikan data ke dalam teknologi *WebGis* menggunakan *teknik information Retrieval*. Menggunakan metode *Vector Space Model* menggunakan metode *cosine similarity*. Dalam pembobotan menggunakan TF-IDF. Data yang digunakan data layanan kesehatan Dokter Praktek(41) + Apotek(37) + Rumah sakit(18) + puskesmas(15) = 111 data, yang dilakukan oleh Subari & Ferdinandus pada tahun 2015. Pada penelitian ini dilakukan untuk mengklasifikasi berita kesehatan termasuk dalam kategori *hoax* atau fakta. Dalam proses pembobotan menggunakan TF-IDF dan *cosine similarity*. Metode yang digunakan *Modified K-Nearest Neighbor* yang dilakukan oleh Prasetyo, Indriati & Adikara pada tahun 2018. Pada penelitian ini dilakukan untuk menerapkan berbagai model *preprocessing* pada analisis sentimen dari teks komentar youtube, kemudian dilihat pengaruhnya pada akurasi model *classifier*. Fitur yang digunakan adalah Unigram dan kombinasi unigram-biagram. Ekstraksi fitur yang digunakan Count-Vectorizer dan TF-IDF-Vectorizer yang dilakukan oleh Khomsah & Aribowo pada tahun 2020. Pada penelitian ini dilakukan untuk menganalisis sentimen terhadap komentar video youtube, untuk memudahkan kreator untuk mengetahui jenis video yang diminati penonton. Ekstraksi fitur yang dilakukan pembobotan setiap kata menggunakan TF-IDF-*Vectorizer*. Metode yang digunakan pada penelitian ini menggunakan *Support Vector Machines* (SVM). media sosial yang digunakan youtube yang dilakukan oleh Muhayat, Fauzi & Indra pada tahun 2023. Terkait dengan penelitian yang sekarang dilakukan untuk menganalisis sentimen masyarakat menggunakan media sosial youtube yang membahasa topik masuk sekolah pukul 5 pagi. Metode yang digunakan menggunakan *K-Nearest Neighbor*, untuk menghitung nilai akurasi menggunakan *confusion matix*, untuk melakukan pengukuran kata dengan menghitung jumlah kemunculan kata pada dokumen menggunakan *Countvectorizer*.

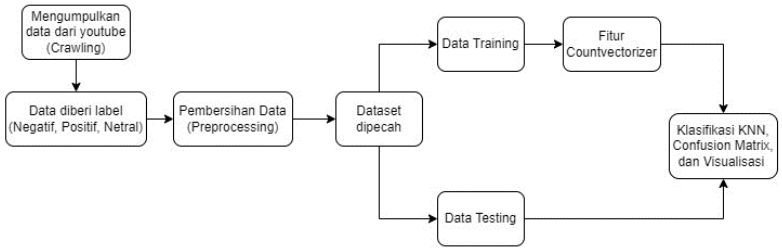
# 2. METODE PENELITIAN

## 2.1 Data Penelitian

Data yang digunakan dalam penelitian ini merupakan dataset yang dikumpulkan dari media sosial *youtube* yang berfokus pada data komentar terkait kebijakan masuk sekolah pukul 5 pagi di NTT yang sedang dibahas di *channel* CNN Indonesia sekitar awal bulan Maret 2023 dengan menggunakan video id yang berbeda dalam satu *channel*. Data yang dipanggil sebanyak 600 data pada program dan data yang berhasil dikumpulkan sebanyak 573 data.

## 2.2 Penerapan Metode

Dalam penelitian ini, beberapa tahapan dijalankan dalam merancang aplikasi/web analisis sentimen. Tahapan ini mewakili setiap proses dan rancangan penelitian dari awal hingga akhir sistem berjalan. Penerapan metode dapat dilihat pada Gambar 1.




**Gambar 1**. Penerapan Metode

Pada Gambar 1 penerapan metode, proses pengumpulan data dilakukan dari komentar youtube. Data yang digunakan membahas tentang kebijakan masuk sekolah pukul 5 pagi di NTT. Selanjutnya proses memberi label secara manual pada dataset yang dilakukan oleh 4 orang responden dari jurusan Bahasa dan Sastra Indonesia, dengan mencari label negatif, positif, dan netral. proses pengambilan labelnya dengan menentukan label terbanyak untuk mencari label hasil akhirnya. Setelah data di beri label masuk ke proses untuk membersihkan data dari kata-kata yang tidak memiliki makna. Terdapat beberapa langkah dari proses preprocessing seperti *case folding*, *cleansing*, *slangword*, *stopword*, *stemming*, *tokenizing*. Selanjutnya data dipecah dengan perbandingan rasio 80:20,

80% untuk data training dan 20% untuk data testing. Selanjutnya tahap *Counvectorizer*, proses ini untuk mencari nilai frekuensi kata yang sering muncul dalam kalimat yang digunakan dengan mengubah fitur teks menjadi sebuah representasi *vector*. Setelah itu tahapan klasifikasi menggunakan *K-Nearest Neighbors* dan menerapkan fitur *Confusion Matrix* untuk mengentahui nilai akurasi, presisi, *recall*, dan *f1-Score*. Tahap terakhir menampilkan hasil visualisasi.

### 2.3 Pengumpulan Data

Proses pengumpulan data komentar youtube dilakukan dengan mengakses *youtube* API yang disediakan oleh *Google* API *Console* dengan mengikuti tahap sampai mendapatkan API *key*, kemudian dilakukan akses dengan memanggil API *key* dan video id pada sistem. Data yang dikumpulkan berasal dari media sosial *youtube*. pada data komentar terkait kebijakan masuk sekolah pukul 5 pagi di NTT yang sedang dibahas di *channel* CNN Indonesia sekitar awal bulan Maret 2023.

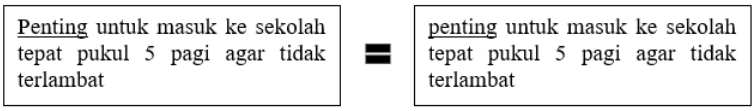
### 2.4 Pelabelan Manual

Pada proses pelabelan untuk memberikan label pada data komentar yang sudah di *crawling*, lalu diberi label dengan ketentuan 0 berlabel negatif, 1 berlabel positif, dan 2 berlabel netral. Proses pelabelan dibantu oleh 4 responden yang berasal dari jurusan Bahasa dan Sastra Indonesia. Kemudian setelah data terkumpul dari masing-masing 4 responden tersebut dicari sentimen terbanyak yang dihasilkan.

### 2.5 *Preprocessing*

Pada proses *preprocessing* untuk membersihkan data dari kata-kata yang tidak diinginkan dan tidak berguna, dengan memanggil fungsi penghilangan simbol, tanda baca, URL, *tag* (*cleansing*), mengubah data dari huruf kapital menjadi huruf kecil (*casefolding*), pemecahan data dari kalimat menjadi kata-kata (*tokenisasi*), mengganti kata-kata yang slang, diganti dengan kata-kata yang formal (*slangword*), pembersihan data dari kata yang tidak penting (*stopword*), menghilangkan kata yang memiliki imbuhan (*stemming*) yang diterapkan ke dataset dalam tampilan *preprocessing*. Berikut adalah beberapa contoh penerapan dari proses *preprocessing* sebagai berikut:

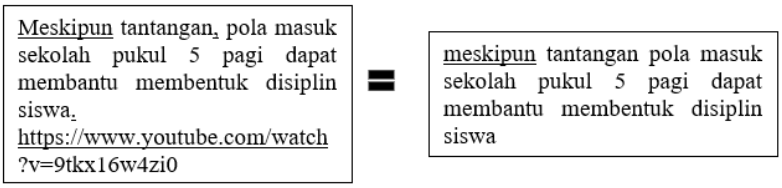
a. *Case Folding*




**Gambar 2** Proses *Case Folding*

Pada gambar 2 merupakan proses *case folding* yang mengonversi semua karakter dalam dokumen dari huruf besar menjadi huruf kecil (*lowercase*).

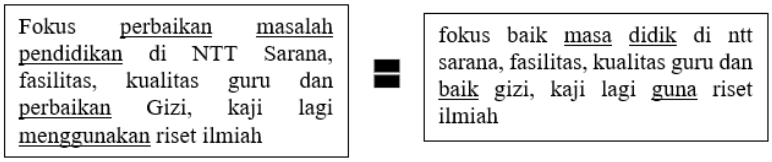
b. *Cleansing*




**Gambar 3** Proses *Cleansing*

Pada gambar 3 merupakan proses *cleansing*, tahap ini dilakukan penghilangan pada simbol, tanda baca, URL.

c. *Stemming*




**Gambar 4** Proses *Stemming*

Pada gambar 4 merupakan proses *stemming*, tahap mengubah kata menjadi kata dasar.

### 2.6 Pembagian Data

Setelah proses data dibeli label dan pembersihan data, kemudian dataset akan dibagi untuk mendapatkan data training dan data testing. Dalam penelitian ini, menggunakan perbandingan 80% untuk data training dan 20% untuk data testing. Pembagian dataset ini dibagi secara random/acak.

### 2.7 *Countvectorizer*

Pada proses *countvectorizer* untuk mencari nilai frekuensi kemunculan kata dalam kalimat dan ditempatkan dalam sebuah vektor pada pemodelan yang akan digunakan untuk proses klasifikasi. Tahapan *CountVectorizer* data dapat dilihat pada contoh berikut ini:

Dokumen 1 : aneh tingkat mutu didik ikat didik mutu

Dokumen 2 : kayak latih militer

Pada contoh data komentar dokumen 1 dan 2 dapat dihitung kemunculan kata pada dokumen. Dapat dilihat pada tabel berikut.

**Tabel 1.** Hasil *Countvectorize*r

| Kata | D1 | D2 |
|---|---|---|
| Aneh | 1 | 0 |
| Tingkat | 1 | 0 |
| mutu | 2 | 0 |
| Didik | 2 | 0 |
| Ikat | 1 | 0 |
| kayak | 0 | 1 |
| Latih | 0 | 1 |
| Militer | 0 | 1 |

Pada Tabel 1 merupakan hasil Countvectorizer untuk dua dokumen, D1 dan D2. Dokumen D1 mencakup kata-kata seperti "Aneh", "Nyata", "Nama", "Tingkat", "mutu", "Didik", dan "Ikat", sementara dokumen D2 berisi kata-kata "kayak", "Latih", dan "Militer".

Dalam dokumen D1, kata-kata "Aneh", "Nyata", "Nama", "Tingkat", "Ikat" muncul satu kali, sementara "mutu" dan "Didik" muncul dua kali. Di sisi lain, dokumen D2 hanya berisi kata "kayak", "Latih", dan "Militer" yang masing-masing muncul satu kali.

### 2.8 *K-Nearest Neighbors*

Pada proses pengujian *K-Nearest Neighbors*, dengan menerapkan fungsi *CountVectorizer* yang terdapat pada proses pemodelan menghasilkan nilai bobot setiap kata yang terdapat pada dokumen. Kemudian nilai bobot tersebut disimpan untuk dilakukan prediksi pada data *testing* untuk diklasifikasi menggunakan algoritme KNN. Dalam algoritme ini, metode yang umum digunakan untuk mengukur jarak perhitungan antar tertangga dokumen terdekat untuk klasifikasi teks adalah *Euclidean Distance*.

$$\textbf{\textit{Euclidean Distance}}\ (\mathbf{A, B}) = \sum_{i=1}^{t} \sqrt{(A - B)^2} \qquad (5)$$

Keterangan :
A : Dokumen *testing* atau data uji
B : Dokumen *training* atau data latih
t : jumlah *term* atau kata

## 3. HASIL DAN PEMBAHASAN

Berdasarkan hasil pembagian dataset ke data *training* dan *testing* yang telah dilakukan proses *preprocessing* sebanyak 573 data. Kemudian akan dilakukan proses pembagian data terhadap dataset menjadi data *training* dan data *testing* dengan perbandingan 80% data *training* dan 20% data *testing*, maka menghasilkan 459 data *training* dan 114 data *testing*.

### 3.1 Tahap pengumpulan Data

Tahap pengumpulan data berasal dari komentar youtube yang membahas tentang kebijakan masuk sekolah pukul 5 pagi, dengan menggunakan dua video id dari channel CNN Indonesia yang dibahas pada channel tersebut sekitar awal bulan maret 2023. Data yang terpanggil sebanyak 600 data dan yang berhasil terkumpulkan sebanyak 573 data.

### 3.2 Tahap Labeling Manual

Tahap pemberian label dilakukan dengan memberikan label pada data komentar dengan nilai 0 untuk label "negatif", nilai 1 untuk label "positif", dan nilai 2 untuk label "netral". Pada proses pelabelan menggunakan 4 responden untuk memberikan label pada komentar kemudian, dihitung hasil label terbanyaknya.

### 3.3 Tahap Preprocessing

Pada tahap *preprocessing*, data dibersihkan dari kata-kata yang tidak diinginkan dann tidak berguna. Dengan menerapkan fungsi *cleansing, casefolding, slangword, stopword,* tokenisasi, *stemming.*

**Tabel 2.** Hasil dari sebelum dan sesudah tahap *Preprocessing*

| Sebelum | Sesudah |
|---|---|
| pak Jokowi tolong di revisi kebijakan gubernur ini pak tolong!!!! Pak Nadiem bantuan pak!!!! | jokowi tolong revisi bijak gubernur tolong nadiem bantu |
| INI BARU CERMINAN STANDARD SERAGAM NEGERI YANG SEBENARNYA,,BUKAN SERAGAM YG DI WAJIBKAN HIJAB DAN ROK PANJANG ????????????? | ini baru cerminan standard seragam negeri yang sebenarnya,,bukan seragam yg di wajibkan hijab dan rok panjang |

Pada tabel 2 hasil dari sebelum dan sesudah tahap preprocessing, pada tabel sesudah merupakan hasil dari pemanggilan keseluruhan fungsi *preprocessing.*

### 3.4 Tahap *Countvectorizer*

Pada tahap *countvectorizer* untuk mencari nilai frekuensi kata yang sering muncul dalam kalimat dan nantinya hasil dari countvectorizer di ubah menjadi representasi vector yang akan digunakan pada klasifikasi *K-nearest Neighbors.*

**Tabel 3.** Komentar bersih pada data Training

| Komentar | Label |
|---|---|
| jam mantap manfat bangun pagi hirup udara pagi | |
| sekolah jalan kaki sepeda sehat tu siswa siswi nya | 1 (Positif) |
| nt mrmasuki dunia kerja rutinitas turut masuk jam | 1 (Positif) |
| disiplin | 0 (Negatif) |
| bilang sana terang timur beda wib | 0 (Negatif) |
| sumpah, gubernur, walikota, nya | |

Pada tabel 3 terdapat beberapa sampel data *training* yang sudah diberi label secara manual dengan 0 label 'negatif', 1 label 'positif'.

**Tabel 4.** Komentar bersih pada data Testing

| Komentar |
|---|
| tri didik indonesia cocok tri molor |
| Teliti bagus |

Pada tabel 4 terdapat beberapa sampel data *testing* yang sudah diberi label secara manual dengan 0 label 'negatif', 1 label 'positif'. Setelah semua data diperoleh, dilakukan perhitungan dengan mengunakan ekstraksi *countvectorizer* untuk mendapatkan nilai frekuensi kemunculan kata yang muncul dalam kalimat. Dapat dilihat pada tabel 5 berikut untuk tahap *countvectorizer*.

**Tabel 5.** Komentar bersih pada data Testing

| Data Training | | | | | | | Testing | |
|---|---|---|---|---|---|---|---|---|
| term | d1 | d2 | d3 | d4 | d5 | | d1 | d2 |
| jam | 1 | 1 | 0 | 0 | 1 | | 0 | 0 |
| mantap | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| manfat | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| bangun | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| pagi | 2 | 0 | 0 | 0 | 0 | | 0 | 1 |
| hirup | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| udara | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| sekolah | 1 | 0 | 0 | 0 | 1 | | 0 | 0 |
| jalan | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| kaki | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| sepeda | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| sehat | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| tu | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| siswa | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| siswi | 1 | 0 | 0 | 0 | 0 | | 0 | 0 |
| nya | 1 | 0 | 0 | 1 | 0 | | 0 | 0 |
| nt | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| mrmasuki | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| dunia | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| kerja | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| rutinitas | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| turut | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| masuk | 0 | 1 | 0 | 0 | 0 | | 0 | 1 |
| disiplin | 0 | 1 | 0 | 0 | 0 | | 0 | 0 |
| bilang | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| sana | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| terang | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| timur | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| beda | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| wib | 0 | 0 | 1 | 0 | 0 | | 0 | 0 |
| sumpah | 0 | 0 | 0 | 1 | 0 | | 0 | 0 |
| gubernur | 0 | 0 | 0 | 1 | 0 | | 0 | 0 |
| walikota | 0 | 0 | 0 | 1 | 0 | | 0 | 0 |
| habis | 0 | 0 | 0 | 0 | 1 | | 0 | 0 |
| pikir | 0 | 0 | 0 | 0 | 1 | | 0 | 0 |
| tri | 0 | 0 | 0 | 0 | 0 | | 2 | 0 |
| didik | 0 | 0 | 0 | 0 | 0 | | 1 | 0 |
| indonesia | 0 | 0 | 0 | 0 | 0 | | 1 | 0 |
| cocok | 0 | 0 | 0 | 0 | 0 | | 1 | 0 |
| molor | 0 | 0 | 0 | 0 | 0 | | 1 | 0 |
| teliti | 0 | 0 | 0 | 0 | 0 | | 0 | 1 |
| bagus | 0 | 0 | 0 | 0 | 0 | | 0 | 1 |

### 3.5 Tahap *K-Nearest Neighbors*

Tahap Klasifikasi *K-Nearest Neighbor* (KNN) Setelah proses mendapatkan nilai frekuensi kemunculan kata dalam sebuah kalimat dengan fitur *countvectorizer*, pada tahap ini hasil dari *countvectorizer* akan dihitung jarak antara sampel data *training* dan data *testing*. Berdasarkan persamaan (1), proses perhitungan jarak antara data *training* dan data *testing* menggunakan rumus *euclidean distance*. Berikut contoh penerapan *euclidean distance* untuk menghitung jarak tetangga terdekatnya pada data testing (A) dan data training (B):

$$(A-B) = \begin{vmatrix} (0-1)^2+(0-1)^2+(0-1)^2+(0-1)^2+ \\ (0-2)^2+(0-1)^2+(0-1)^2+(0-1)^2+ \\ (0-1)^2+(0-1)^2+(0-1)^2+(0-1)^2+ \\ (0-1)^2+(0-1)^2+(0-1)^2+(0-1)^2+ \\ (0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+ \\ (0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+ \\ (0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+ \\ (0-0)^2+(0-0)^2+(0-0)^2+(0-0)^2+ \\ (0-0)^2+(0-0)^2+(0-0)^2+(2-0)^2+ \\ (1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+ \\ \sqrt{(0-0)^2+(0-0)^2} \end{vmatrix}$$

$$(A-B) = \begin{vmatrix} 1+1+1+1+4+ \\ 1+1+1+1+1+1+ \\ 1+1+1+1+1+1+ \\ 1+0+0+0+0+0+ \\ 0+0+0+0+0+0+ \\ 0+0+0+0+0+0+ \\ 0+0+0+0+0+0+ \\ 4+1+1+1+1+1+ \\ \sqrt{0+0} \end{vmatrix}$$

$$(A-B) = \sqrt{27}$$

$$(A-B) = \sqrt{5{,}196152423}$$

Berdasarkan hasil dari perhitungan *euclidean distance*, maka penentuan jarak antara data *training* dan data *testing* di urutkan dari yang terkecil ke terbesar. Dapat dilihat hasil perhitungan *euclidean* dalam menentukan jarak antara data *training* dan data *testing* pada tabel 6 berikut.

**Tabel 6.** Tabel Hasil Euclidean Distance

| | Data Testing 1 | Label |
|---|---|---|
| Data *Training* 4 | 3,464101615 | 0 (Negatif) |
| Data *Training* 5 | 3,464101615 | 0 (Negatif) |
| Data *Training* 3 | 3,741657387 | 0 (Negatif) |
| Data *Training* 2 | 4,123105626 | 1 (Positif) |
| Data *Training* 1 | 5,196152423 | 1 (Positif) |

Berdasarkan tabel 6 dihitung label yang sering muncul dalam tetangga terdekat menghasilkan prediksi klasifikasi. Dan dapat disimpulkan bahwa dari hasil perhitungan jarak menggunakan rumus *Euclidean Distance* dan menggunakan data *training* 5 sampel dan data *testing* 2 sampel, maka hasil klasifikasi bersentimen negatif.

### 3.6 Pengujian

Pada tahap pengujian dalam penelitian ini membuktikan sistem yang diimplementasikan sudah sesuai dengan ketentuan dan rancangan yang sudah direncanakan sebelumnya. Pada pengujian Confusion Matrix. Berdasarkan hasil prediksi dari pemodelan data training dan data testing. Data terdiri dari 459 data training dan 114 data testing. Hasil dari data testing menhasilkan sentimen 105 untuk negatif, 9 untuk positif, dan 0 untuk netral.berikut tabel 7 hasil klasifikasi algoritme *K-Nearest Neighbors.*

**Tabel 7.** Hasil klasifikasi algoritme *K-Nearest Neighbors*

| No | Komentar | Label Aktual | Label Prediksi |
|---|---|---|---|
| 1 | Gubernur ntt cuma membuat sensasi saja, cuma mencari masalah dengan menambah masalah. Kapan asn masuk kerja jam 5.30 | Negatif | Negatif |

| | | | |
|---|---|---|---|
| 2 | Ayo pak menteri pendidikan, bersuara lah ini kebijakan gubernur yang nabrak aturan, kebijakan yang menyusahkan siswa dan orang tua nya | Negatif | Netral |
| 3 | Namanya juga sekolah indo, selalu ngajarin hal hal yg gk sesuai sama kondisi masyarakat | Negatif | Negatif |
| ----- | ---------- | ---------- | ---------- |
| 114 | Belajar di sekolah menurut saya tidak harus lama tetapi bs ditambah ektrakurikuler wajib dan tidak wajib | Netral | Positif |

Pada tabel 5 Tabel Hasil klasifikasi *K-Nearest Neighbors*. Pada kolom label aktual hasil yang diperoleh dari proses manual labeling, sementara label prediksi hasil yang yang diperoleh dari klasifikasi menggunakan algoritme KNN. Untuk keseluruhan data testing yang terdapat di proses pengujian terdapat 114 data. kemudian direpresentasi mengunakan *confusion matrix*. Pada pengujian *confusion matrix* merepresentasikan tetangga terdekatnya terbentuk dapat dilihat pada tabel 8 berikut.

**Tabel 8.** *Confusion Matrix* Pengujian

| | | Nilai Prediksi | | |
|---|---|---|---|---|
| | | Positif | Negatif | Netral |
| Nilai Aktual | Positif | TP 6 | FP 2 | Fnet 12 |
| | Negatif | Fnet 12 | Tneg 78 | Fneg 16 |
| | Netral | Fnet 12 | Fneg 16 | Tnet 0 |

Berdasarkan tabel 8 *Confusion Matrix* pengujian, hasil dari perolehan tabel *confusion matrix* mendapatkan jumlah sentimen aktual dan prediksi, mendapatkan hasil *True* Positif (TP) 6, *False* Positif (FP) 2, *False* Netral (Fnet) 12, *True* Negatif (Tneg) 78, *False* Negatif (Fneg) 16, *True* Netral (Tnet) 0. maka perolehan dari tabel *confusion matrix* kemudian dicari nilai akurasi, presisi, *recall*, *f1-score* dapat dilihat pada tabel 9. dibawah ini.

**Tabel 9.** Nilai Pengujian

| **Pengujian** | | |
|---|---|---|
| Akurasi | $= \frac{6+78+0}{6+2+78+16+0+12}\ x\ 100$ | 0.7368 (73.68%) |
| Presisi | $= \frac{6}{6+2}$ x 100 | 0.75 (75%) |
| Recall | $= \frac{6}{6+16+12}$ x 100 | 0.1764 (17.65%) |
| F1-Score | $= 2.\frac{75\ x\ 17.65}{75+17.65}\ x\ 100$ | 14.28 (28.58%) |

Pada tabel 9 merupakan hasil perhitungan yang dilakukan secara manual berdasarkan keluaran nilai dari *confusion matrix*. Pengujian diatas dilakukan secara berulang dengan variasi nilai yang berbeda-beda. Sehingga dapat diketahui hasil pengujian secara keseluruhan seperti pada tabel 10 dibawah ini.

**Tabel 10.** Hasil Pengujian Tetangga Terdekat

| | K=3 | K=5 | K=7 | K=9 | K=11 |
|---|---|---|---|---|---|
| Akurasi | 71% | 70% | 71% | 72% | 73% |
| Presisi | 50% | 50% | 53% | 63% | 75% |
| *Recall* | 25% | 23% | 24% | 20% | 17% |
| F1-Score | 33% | 32% | 33% | 31% | 28% |

Berdasarkan tabel 4.8 diatas diketahui dengan nilai akurasi tertinggi didapatkan dengan nilai k=11, yaitu akurasi 73%, presisi 75%, recall 17%, dan f1-score 28%. Sedangkan nilai terendah didapatkan dengan nilai k=5, yaitu akurasi 70%, presisi 50%, recall 23% dan f1-score 32%. Dari hasil penelitian tersebut bisa disimpulkan

bahwa nilai K bersifat tidak menentu, harus dicari dan bergantung pada jenis, dan bobot yang digunakan dalam proses pengujian.

## 4. KESIMPULAN

Berdasarkan hasil evaluasi dari aplikasi untuk menganalisis sentimen masyarakat terhadap kebijakan masuk sekolah pukul 5 pagi di NTT dengan menggunakan data komentar youtube, maka dapat disimpulkan bahwa, berdasarkan sistem aplikasi analisis sentimen cenderung ke arah sentimen negatif. sistem dapat menganalisis sentimen masyarakat dalam kategori negatif, positif, netral, dan pada pengujian mendapatkan hasil pengujian dengan tetangga terdekatnya k=11 dengan nilai pengujian tertinggi diperoleh dengan nilai akurasi 73.68%, presisi 75%, *recall* 17.65%, *f1-score* 28.58%. adapun saran yang dapat peneliti berikan sebagai pengembangan lebih lanjut untuk sistem ini agar dapat berjalan dengan lebih baik lagi menggunakan data yang lebih akurat, memerlukan dataset yang lebih banyak dan terkini, menggunakan channel yang berbeda dalam proses pengambilan dataset agar tau perbedaan sentimen lebih banyak ke arah negatif, positif atau netral.

## DAFTAR PUSTAKA

[1] Hermansyah, "Manajemen Lembaga Pendidikan Sekolah Berbasis Digitalisasi Di Era Covid 19," *Fitrah : Jurnal Studi Pendidikan*, vol. 12, no. 1, pp. 28–46, 2021, doi: https://doi.org/10.47625/fitrah.v12i1.320.

[2] M. P. R. Putra and K. R. N. Wardani, "PENERAPAN TEXT MINING DALAM MENGANALISIS KEPRIBADIAN PENGGUNA MEDIA SOSIAL," *JUTIM (Jurnal Teknik Informatika Musirawas)*, vol. 05 no 1, pp. 63–71, 2020, [Online]. Available: http://eprints.binadarma.ac.id/11221/

[3] Isman, A. Ahmad, and A. Latief, "Perbandingan Metode KNN Dan LBPH Pada Klasifikasi Daun Herbal," *Jurnal Resti (Rekayasa Sistem dan Teknologi Informasi)*, vol. 1, no. 10, pp. 557–564, 2021, doi: https://doi.org/10.29207/resti.v5i3.3006.

[4] F. Subari, "SISTEM INFORMATION RETRIEVAL LAYANAN KESEHATAN UNTUK BEROBAT DENGAN METODE VECTOR SPACE MODEL (VSM) BERBASIS WEBGIS," *Seminar Nasional Teknologi Informasi, Komunikasi dan Aplikasinya*, vol. 03, pp. 202–212, 2015.

[5] A. Prasetyo, Rino, Indriati, and P. Adikara, Pandu, "Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 12, pp. 7466–7473, 2018, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/3991

[6] S. Khomsah and Agus Sasmito Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 648–654, 2020, doi: 10.29207/resti.v4i4.2035.

[7] F. A. Nugraha, N. H. Harani, and R. Habibi, *ANALISIS SENTIMEN TERHADAP PEMBATASAN SOSIAL MENGGUNAKAN DEEP LEARNING*. Bandung: Kreatif Industri Nusantara, 2020. [Online]. Available: https://books.google.co.id/books?id=f738DwAAQBAJ&printsec=frontcover&hl=id#v=onepage&q&f=false

[8] T. Muhayat, A. Fauzi, and D. J. Indra, "Analisis Sentimen Terhadap Komentar Video Youtube Menggunakan Support Vector Machines," *Progresif: Jurnal Ilmiah Komputer*, vol. 19, pp. 231–240, 2023, [Online]. Available: http://ojs.stmik-banjarbaru.ac.id/index.php/progresif/article/view/1060

[9] 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," *The 6th International Conference on Cyber and IT Service Management (CITSM 2018)*, no. June 2021, pp. 1–6, 2018, doi: 10.1109/CITSM.2018.8674352.

[10] D. A. Muthia, "ANALISIS SENTIMEN PADA REVIEW BUKU MENGGUNAKAN ALGORITMA NAÏVE BAYES," *Jurnal Paradigma*, vol. XVI, no. 1, p. 12, 2014, doi: 10.31294/p.v16i1.723.