**ESTIMASI: Journal of Statistics and Its Application**
Vol 4, No. 2, Juli, 2023, Hal. 196-207
DOI: https://doi.org/10.20956/ejsa.v4i2.25522

# Analisis Sentimen Survei Regsosek pada *Twitter* Menggunakan Algoritma *K-Nearest Neighbor* (K-NN)

**Bunga Ayuningrum[1], Hilma Hanna Mahanna Haqq[2], Suci Mega Puji Lestari[3], M Al Haris[4]**
**[1,2,3,4]Program Studi Statistika, Fakultas MIPA,**
**Universitas Muhammadiyah Semarang, Semarang, 50273, Indonesia**
*** Corresponding author, email: bungaaya04@gmail.com**

***Abstract***

*Indonesia in 2022, will experience a shift in adaptation to recovery from the pandemic as well as rising global commodity prices due to the impact of the Ukraine-Russia war. The government in its efforts to deal with this situation, one of which is by transforming data into one data through the 2022 Social Economic Registration (Regsosek) as a requirement for social protection system reform. However, in practice, Research and Research has become quite a public concern, where the content is almost the same as previous surveys conducted by BPS, which raises questions about the effectiveness of this survey. This study aims to determine the sentiments of each opinion on social media Twitter regarding 2022 Social Security. This research implements the K-Nearest Neighbor (K-NN) method to analyze sentiment in tweets. Data obtained from Twitter by scrapping. The polarity percentage results from the tweets obtained are dominated by negative opinions. The best application of the K-Nearest Neighbor (K-NN) algorithm is using the parameter k = 3. The model built shows very good performance with an accuracy of 96%, a recall of 100%, and a precision of 0,96%.*

***Keywords****: K-Nearest Neighbor, Regsosek, Sentimen analysis, Twitter*

**Abstrak**

Indonesia pada tahun 2022, mengalami pergeseran adaptasi pemulihan dari pandemi serta kenaikan harga komoditas global akibat dampak perang Ukraina-Rusia. Pemerintah dalam upayanya untuk menghadapi situasi tersebut, salah satunya adalah dengan melakukan transformasi data menjadi satu data melalui Registrasi Sosial Ekonomi (Regsosek) 2022 sebagai persyaratan reformasi sistem perlindungan sosial. Namun dalam pelaksanaannya, Regsosek cukup menjadi perhatian publik, dimana isinya hampir sama dengan survei-survei sebelumnya yang telah dilakukan oleh BPS sehingga menimbulkan pertanyaan mengenai keefektifan dari survei ini. Penelitian ini bertujuan untuk mengetahui sentimen dari setiap opini di sosial media *Twitter* mengenai Regsosek 2022. Penelitian ini mengimplementasikan metode *K-Nearest Neighbor (K-NN)* untuk menganalisa sentimen pada *tweets*. Data diperoleh dari *twitter* dengan cara *scrapping*. Hasil presentase polaritas dari *tweets* yang didapat, didominasi oleh opini negatif. Penerapan algoritma *K-Nearest Neighbor (K-NN)* terbaik adalah menggunakan parameter k= 3. Model yang dibangun menunjukan performa yang sangat bagus dengan akurasi sebesar 96%, *recall* 100%, dan *precision* 0,96%.

**Kata Kunci**: *K-Nearest Neighbor*, Regsosek, Analisis Sentimen, *Twitter*

## 1. Pendahuluan

Menurut Badan Pusat Statistik, pandemi *covid*-19 menyebabkan perekonomian Indonesia menurun pada tahun 2020-2021 dengan ditandainya tingkat pengangguran dan kemiskinan lebih tinggi dari tahun-tahun sebelumnya. Meskipun pada tahun 2022 tingkat pengangguran dan kemiskinan lebih rendah dibandingkan tahun 2021, namun angka

tersebut masih lebih tinggi dibandingkan masa sebelum *covid*-19. Pada tahun 2022 pergeseran adaptasi pemulihan pandemi serta kondisi kenaikan harga komoditas global akibat dampak perang Ukraina-Rusia menjadi peluang peningkatan kemiskinan di Indonesia. Guna menghadapi situasi tersebut pemerintah mengeluarkan tiga perubahan struktural, salah satunya adalah sistem perlindungan sosial [1].

Transformasi data menjadi satu data melalui Regstrasi Sosial Ekonomi (Regsosek) 2022 menjadi salah satu syarat utama terpenuhinya transformasi tersebut. Regsosek 2022 bertujuan untuk menyediakan kumpulan data populasi tunggal yang akan memungkinkan pemerintah untuk mengimplementasikan berbagai program dengan terkoordinasi, tidak tumpang tindih, dan lebih efektif [2]. Namun, Regsosek cukup menjadi perhatian publik dalam pelaksanaannya, dimana isinya yang hampir sama dengan survei atau sensus sebelumnya, seperti Survei Sosial Ekonomi Nasional (Susenas), Sensus Penduduk Lanjutan, dan lain-lain, menimbulkan pertanyaan tentang keefektifan survei ini. Kondisi inilah yang menarik peneliti untuk melakukan analisis sentimen mengenai Survei Regsosek 2022 pada media sosial *twitter* dengan mengklasifikasikan secara otomatis opini negatif atau positif masyarakat terhadap survei ini.

Analisis sentimen pada *twitter* sudah banyak dilakukan oleh beberapa peneliti terdahulu, diantaranya Fauziyyah dan Gautama (2020) melakukan analisis sentimen mengenai pandemi *covid*-19 pada *streaming twitter* dengan menggunakan *text mining* [3]. Namun, pada penelitian tersebut peneliti tidak menyebutkan algoritma yang digunakan pada *TextBlob* dan bagaimana performa dari model tersebut. Rofiqoh, Perdana, dan Fauzi (2017) juga mengembangkan analisis sentimen untuk tingkat kepuasan pengguna layanan telekomunikasi seluler Indonesia pada *twitter* [4]. Penelitian ini mengembangkan sentimen analisis dengan menggunakan metode *Support Vector Machine* dan *Lexicon Based Features*, dengan akurasi yang dihasilkan sebesar 79%. Analisis sentimen juga dimplementasikan pada penelitian Somantri dan Dairoh (2019) untuk penilaian tempat tujuan wisata kota Tegal [5]. Pada penelitian ini, peneliti menerapkan dua metode, yaitu *Naïve Bayes* dan *Decision Tree* dengan akurasi yang dihasilkan untuk metode *naïve bayes* sebesar 77,50% dan *Decision Tree* sebesar 60,83%. Penelitian Furqan, Sriani, dan Sari (2022) juga melakukan analisis sentimen terhadap kasus *new normal* masa *covid*-19 dengan menerapkan metode *K-Nearest Neighbor (K-NN).* Pada penelitian ini, metode *K-Nearest Neighbor (K-NN)* menghasilkan akurasi yang sangat tinggi yaitu sebesar 100%. [6].

Berdasarkan permasalahan di atas, maka peneliti berkeinginan untuk mengimplementasikan analisis sentimen pada *twitter* pada topik permasalahan Survei Regsosek 2022. Penelitian ini akan menerapkan algoritma *K-Nearest Neighbor (K-NN),* dimana pada penelitian sebelumnya algoritma ini menghasilkan akurasi yang tinggi.

## 2. Metode dan Analisis

### 2.1 Pengumpulan dan Pelabelan Data

Penelitian ini menggunakan data yang diperoleh dari *twitter* dengan cara *scrapping*. Metode pengumpulan data penelitian ini menggunakan *library snscrape* dengan kata kunci yang menjadi *keyword* adalah Regsosek 2022. Data diambil dari tanggal 1 September 2022 sampai 30 November 2022 dengan total *tweets* yang diambil adalah 500 *tweets.*

Data yang berhasil dikumpulkan akan diberi label dengan dua kategori kelas sentimen yaitu positif dan negatif. Kriteria yang digunakan dalam mengkategorikan kelas sentimen adalah *tweets* akan dianggap positif apabila mengandung kata tidak memihak dan dimaksudkan untuk menyetujui adanya survei Regsosek 2022. Namun, *tweets* akan dikategorikan negatif apabila mengandung bahasa tidak baik dan menyatakan ketidaksetujuan terhadap adanya survei ini.

### 2.2 *Preprocessing*

Tahapan selanjutnya setelah data terlabeli adalah *preprocessing,* dimana proses ini membantu menyiapkan data dari teks mentah hingga siap digunakan sesuai dengan metode yang dipakai [7]. Tujuan dari *preprocessing* yaitu untuk meningkatkan akurasi dari model klasifikasi [8]. Terdapat beberapa tahap untuk melakukan *preprocessing* pada *text mining,* yaitu *cleansing, case folding, word normalization, filtering/stopwords removal, stemming*, dan *tokenization* [9].

1. *Cleansing*, merupakan proses menghilangkan huruf asing dan tanda baca dari teks. Tujuan *cleansing* pada tahap ini digunakan untuk mengurangi gangguan dan *noise* dalam kumpulan data. Menghapus URL, pengguna, RT, dan tagar adalah bagian dari tahapan dalam proses *cleansing*. Pada proses ini digunakan salah satu *library* NLTK yaitu *WordNetLemmatizer.*
2. *Case Folding,* merupakan proses yang digunakan untuk mengkonversi teks ke dalam format huruf kecil *(lower case)* atau huruf besar *(upper case),* dimana pada penelitian ini tipe *case folding* yang akan digunakan adalah *lower case*. Tujuan dari proses ini adalah untuk memberikan format atau bentuk yang seragam pada data.
3. *Word Normalization,* merupakan proses untuk menormalisasi atau memberi standar kata yang memiliki makna sama namun penulisannya berbeda, baik diakibatkan kesalahan penulisan, bentuk penyingkatan kata maupun bahasa gaul.
4. *Filtering/stopwords removal,* metode memilih kata-kata penting dalam teks untuk digunakan. Terdapat dua teknik yang dapat diterapkan, yaitu *stoplist* dan *wordlist.* Teknik *stoplist* akan diterapkan pada penelitian ini.
5. *Stemming/root-finding,* merupakan salah satu proses untuk mengubah bentuk kata yang kompleks menjadi kata yang lebih sederhana atau kata dasar. Beberapa teknik dapa diimplementasikan untuk mengubah kata dalam bentuk kata dasar, diantaranya *porter stemmer, stemming nazief-adriani, stemming Arifin-setiono, dan khoja*. Pada penelitian *library* yang digunakan untuk proses *stemming* adalah *library sastrawi* (*Stemming Nazief-Adriani).*

6. *Tokenization,* merupakan proses memisahkan teks menjadi bagian-bagian kecil atau kata-kata yang disebut dengan token. Tahap ini akan menghapus karakter seperti angka, tanda baca dan lainnya yang dianggap berdampak minimal pada pemrosesan.

**2.3 *Feature Extraction* dengan TF-IDF**

Proses yang terjadi pada tahapan ini adalah pembobotan kata, dimana akan dilakukan pemberikan skor/bobot pada jumlah frekuensi kemunculan kata yang ada dalam data. Pada proses ini kemunculan kata akan diberi bobot atau skor dengan menggunakan TF-IDF. *Term Frequency-Invers Document Frequency* atau disebut TF-IDF merupakan teknik pembobotan yang berasal dari dua nilai, dimana nilai tersebut diperoleh dari dua algoritma berbeda. Algorima ini akan menentukan frekuensi relatif dari sebuah kata pada kumpulan data. Frekuensi kemunculan kata *(tf)* dalam data tersebut dapat dihitung dengan persamaan berikut [10].

$$tf\ (i) = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (1)$$

Sedangkan persamaan idf dapat dirumuskan sebagai berikut.

$$idf_i = \frac{log\ |\ D\ |}{|\{d: t_i \epsilon\ d\}|} \quad (2)$$

**2.4 *Feature Selection***

Pada tahap ini, proses yang terjadi adalah mengkategorikan data ke dalam kelas tertentu dengan cara menemukan bentuk polanya yang relevan. Tujuan dilakukannya proses ini adalah untuk memprioritaskan fitur – fitur penting pada selama pemodelan berlangsung. Terdapat beberapa metode yang bisa digunakan pada tahap *feature selection.* Penelitian ini menggunakan metode *Chi Square*, yang dapat dirumuskan dalam persamaan berikut [11].

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

Dengan, nilai O mewakili jumlah munculnya nilai suatu fitur pada tiap kelas. Sedangkan, nilai $E_i$ adalah nilai jumlah kemunculan fitur pada semua kelas yang dapat dihitung dengan persamaan berikut [12].

$$E_i = \frac{m * n}{D} \quad (4)$$

Dengan, m adalah nilai jumlah semua kelas, n sebagai nilai jumlah tiap kelas, dan D adalah jumlah data.

**2.5 Pembagian Data**

Tahap ini merupakan proses memisahkan data menjadi data *training* dan data *testing*. Data latih atau data *training* merupakan data yang digunakan selama proses pelatihan, sedangkan data yang digunakan untuk menguji model hasil pelatihan adalah data uji atau data *testing.* Pada penelitian ini pembagian data menggunakan *test size* 0.2, yang artinya perbandingan antara data latih dan data uji adalah 80:20 dengan 80% untuk data latih dan 20% sisanya untuk data uji.

**2.6 Analisis *K-Nearest Neighbor (K-NN)***

Cara kerja dari algoritma *K-Nearest Neighbor* adalah mengklasifikasikan objek yang paling dekat dengan objek lainnya berdasarkan data latih dengan mencari jarak terdekatnya [13]. Perhitungan jarak pada algoritma ini dapat menggunakan beberapa cara, seperti Jarak *Euclidean, Mahalanobis, Manhattan, Minkowski, Chebyshev, Cosinus, Korelasi, Hamming, Jaccard, dan Spearman* [14]. Penelitian ini menggunakan jarak *Euclidean* untuk mencari jarak terdekat dua tetangga k, dikarenakan teknik ini memiliki implementasi yang mudah, efisiensi dan produktivitas yang baik[15]. Jarak *Euclidean* dapat didefinisikan sebagai berikut [16].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_{training}^{i} - y_{testing}^{i})^2} \quad (5)$$

Dengan,

*d(x,y)* = jarak *Euclidean Distance,*
$x_{training}^{i}$ = data *training*,
$y_{testing}^{i}$ = data *testing*,
*i* = variabel data,
n = dimensi data

**2.7 Evaluasi**

Tahap evaluasi digunakan untuk menilai seberapa baik model untuk melakukan klasifikasi data. Salah satu metode evaluasi model yaitu *Confusion matrix,* dimana metode ini sudah banyak digunakan dalam beberapa penelitian untuk mengevaluasi hasil dan mengukur kinerja klasifikasi. *Confusion matrix* dapat didefinisikan sebagai berikut [17].

Tabel 1. *Confusion Matrix*

| **Aktual** | **Prediksi** | |
|---|---|---|
| | ***True*** | ***False*** |
| *True* | TP | FP |
| *False* | FN | TN |

Dengan,

| | | |
|---|---|---|
| *True Positive* (TP) | : | Jumlah data yang Positif dan diprediksi benar sebagai data Positif. |
| *False Positive* (FP) | : | Jumlah data yang bernilai Negatif tetapi diprediksi sebagai Positif. |
| *False Negative* (FN) | : | Jumlah data yang bernilai Positif tetapi diprediksi sebagai Negatif. |
| *True Negative* (TN) | : | Jumlah data yang bernilai Negatif dan diprediksi benar sebagai Negatif. |

*Confusion matrix* dalam mengukur kinerja suatu model menggunakan pendekatan *accuracy, precision dan recall* [16].

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

## 3. Hasil dan Diskusi

Berikut merupakan hasil dari tahapan-tahapan yang dilakukan untuk melakukan analisis sentimen mengenai Regsosek pada *twitter.*

### 3.2 Pengumpulan dan Pelabelan Data

Data yang digunakan berjumlah 500 data *tweets* hasil *scrapping* dari *twitter*, yang kemudian dilabeli dengan dua kategori sentimen yaitu negatif dan positif. Berikut merupakan hasil *labeling* sentimen positif dan negatif.

Tabel 2. Hasil *labeling* sentimen

| ***Tweets*** | ***Sentimen*** |
|---|---|
| regsosek mengacaukan konsep ruta sakernas @ krsnd21 | *Negatif* |
| Terimakasih Regsosek dan Event sepeda yg sudah merubah segalanya. @ muhtarsa | *Positif* |

### 3.3 *Preprocessing*

Tahap pertama dalam *preprocessing* adalah proses *cleansing*, dimana pada proses *cleansing* ini dilakukan untuk membersihkan karakter tertentu dari data mentah yang tidak diperlukan seperti *hastag*, alamat *website*, *username*, angka, emoji serta *email*. Berikut merupakan hasil dari proses *cleansing.*

Tabel 3. Hasil *Cleansing*

| **Data Mentah** | ***Cleansing*** |
|---|---|
| @worksfess Gw Cuma sekali ngerasain gaji umr yaitu pas ikut regsosek. Itupun kudu nangis2 dlu menjalani hari hari penuh emosi ketemu responden yg macem-macem bentuknya 😂 | Gw Cuma sekali ngerasain gaji umr yaitu pas ikut regsosek Itupun kudu nangis dlu menjalani hari hari penuh emosi ketemu responden yg macem macem bentuknya |

Selanjutnya, setelah *cleansing* dilakukan tahap berikutnya adalah proses *case folding, word normalization, filtering, stemming, tokenizing.* Tabel 4 merepresentasikan output yang dihasilkan dari tahap-tahap *preprocessing* tersebut.

Tabel 4. Hasil *preprocessing*

| ***Case Folding*** | ***Word Normalization*** | ***Filtering*** | ***Stemming*** | ***Tokenizing*** |
|---|---|---|---|---|
| gw cuma sekali ngerasain gaji umr yaitu pas ikut regsosek itupun kudu nangis dlu menjalani hari hari penuh emosi ketemu responden yg macem macem bentuknya | saya hanya sekali merasakan gaji umr yaitu saat ikut regsosek itupun harus nangis dulu menjalani hari hari penuh emosi ketemu responen yang macam macam bentuknya | merasakan gaji umr saat regsosek itupun harus nangis dulu menjalani penuh emosi ketemu responden macem macem bentuknya | rasa gaji umr saat regsosek itupun harus nangis dulu jalan penuh emosi temu responden macem macem bentuk | rasa<br>gaji<br>umr<br>saat<br>regsosek<br>itupun<br>harus<br>nangis<br>dulu<br>jalan<br>penuh<br>emosi<br>temu |

### 3.4 *Feature Extraction* dengan TF-IDF

Pembobotan kata pada tahap ini menggunakan algoritma TF-IDF (*Term Frequency-Invers Document Frequency)*. Berikut merupakan simulasi dari perhitungan bobot dengan algoritma TF-IDF.

Tabel 5. Penentuan Idf

| **token** | **TF** | | | **DF** | **IDF (log d/df)** |
|---|---|---|---|---|---|
| | **D1** | **D2** | **D3** | | |

| | | | | | |
|---|---|---|---|---|---|
| rasa | 1 | 1 | 0 | 2 | 1,287682 |
| gaji | 0 | 2 | 0 | 1 | 1,693147 |
| umr | 0 | 0 | 1 | 1 | 1,693147 |
| regsosek | 0 | 0 | 1 | 1 | 1,693147 |

Tabel 6. Penentuan bobot

| **BOBOT (TF * IDF)** | | |
|---|---|---|
| Hasil Akhir Normalisasi | | |
| D1 | D2 | D3 |
| 0,428046 | 0,306504 | 0 |
| 0 | 0,806032 | 0 |
| 0 | 0 | 0,377964 |
| 0 | 0 | 0,377964 |

### 3.5 *Feature Selection*

Pada tahap ini proses yang dilakukan adalah mengambil fitur penting yang berguna pada tahap pemodelan *machine learning.* Berikut merupakan nilai *chi square* beserta fitur yang terpilih, dimana jumlah fitur yang diambil sebanyak 200 dari 1831.

Tabel 7. Hasil dari proses *feature selection*

| **Nilai** | **Fitur** |
|---|---|
| 54,206199 | smart |
| 41,51076 | tangan |
| 39,399711 | kampung |
| ⋮ | ⋮ |
| 15.075811 | miskin |

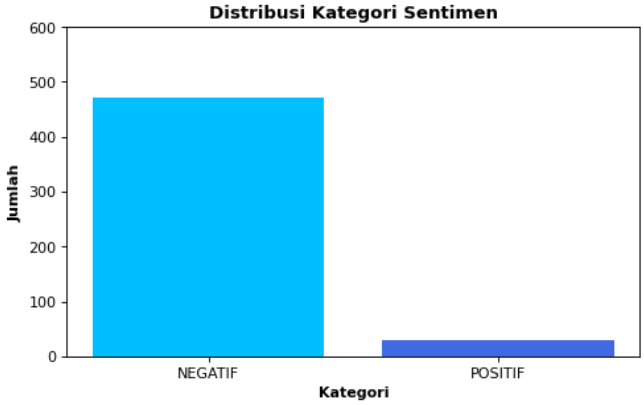
### 3.6 Pemodelan *K-NN*

Pada penelitian ini algoritma *K-Nearest Neighbor (K-NN)* diimplementasikan pada sentimen opini masyarakat mengenai survei Regsosek 2022 di media sosial *twitter* dengan tujuan untuk megetahui seberapa besar tingkat akurasi, *recall dan precision* dari analisis sentimen yang dilakukan. Komputasi *model K-Nearest Neighbor (K-NN)* terhadap bobot tiap kata dalam sentimen sangat menentukan kelas negatif dan positif, serta akurasi yang dihasilkan dari proses pemodelan. Percobaan dilakukan dengan menerapkan parameter nilai k, yaitu k= 3,7,5 guna mendapatkan model terbaik. Berikut merupakan *output* percobaan dari penerapan parameter nilai k.

Tabel 8. Hasil percobaan dengan nilai k

| k | Prediksi Salah | Prediksi Benar | *Accuracy* |
|---|---|---|---|
| 3 | 4 | 97 | 96,04% |
| 5 | 4 | 97 | 95,03% |
| 7 | 4 | 97 | 94,03% |

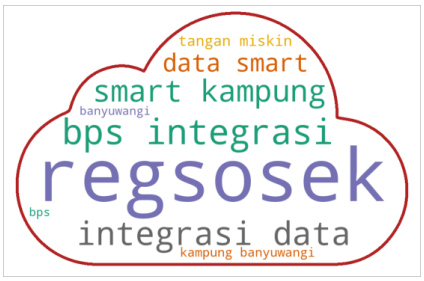
Berdasarkan tabel percobaan diatas menunjukan bahwa penerapan algoritma *K-Nearest Neighbor* (K-NN) terbaik adalah dengan menggunakan parameter k= 3 dengan hasil akurasi yang didapatkan sebesar 96,04%. Berikut merupakan hasil presentase sentimen opini masyarakat mengenai Regsosek 2022 dari tanggal 1 September 2022 sampai 30 November 2022.



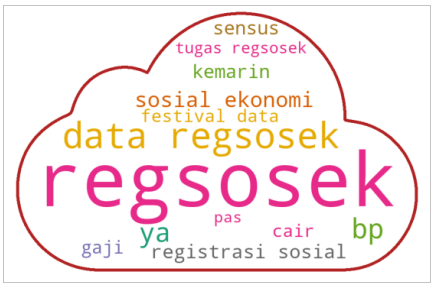

Gambar 1. Polaritas *tweets*

Gambar 1 merupakan hasil representasi polaritas dari *tweets* yang didapat. Dari diagram diatas menunjukkan bahwa opini negatif lebih tinggi dibandingkan opini positif. Dari hasil tersebut, dapat dikatakan bahwa opini masyarakat terhadap survei Regsosek 2022 didominasi oleh opini negatif.

Selanjutnya, kata dominan yang sering muncul dalam kelas positif dan negatif disajikan dalam *word cloud* pada masing-masing kelas berikut. Kata kelas positif ditunjukan pada *word cloud* gambar 2, sedangkan untuk kata kelas negatif ditunjukkan pada gambar 3.




Gambar 2. Kelas positif




Gambar 3. Kelas negatif

Dari kedua *word cloud* diatas menunjukan bahwa pada kelas positif kata yang sering muncul atau sering digunakan masyrakat untuk menyampaikan opininya adalah Regsosek, Integrasi data, data *smart*. Sedangkan untuk kelas negatif, kata yang sering digunakan masyarakat untuk meyampaikan opininya adalah Regsosek, data Regsosek, tugas Regsosek dan sensus sosial ekonomi.

### 3.7 Evaluasi

Pengujian data uji menggunakan model *classifier K-Nearest Neighbor* dengan k yang digunakan berdasarkan hasil percobaan adalah k=3. Model yang telah dibangun tersebut digunakan untuk mengklasifikasikan sentimen kedalam dua kategori, yaitu sentimen positif dan negatif. Dalam penelitian ini, data yang dievaluasi berjumlah 500 data *tweets* dengan topik Regsosek 2022. *Confusion matrix* digunakan untuk mengevaluasi hasil kinerja model *K-Nearest Neighbor* yang telah dirancang. Tabel 9 menunjukkan hasil *confusion matrix* model.

Tabel 9. *Confusion Matrix*

| **Data Aktual** | **Data Prediksi** | |
|---|---|---|
| | **Positif** | **Negatif** |
| Positif | 0 | 0 |
| Negatif | 4 | 97 |

Tabel 10. Performa Model

| **Precision** | **Recall** | **Accuracy** |
|---|---|---|
| 0,96 | 1,00 | 96% |

Berdasarkan tabel *confusion matrix* diatas, menunjukkan bahwa model *classifier* yang dibangun dapat mengidentifikasi 97 sentimen dengan tepat dari 101 data uji. Pada tabel 10, model *K-Nearest Neighbor* yang dibangun menunjukan performa yang sangat bagus dengan akurasi sebesar 96%, *recall* 100%, dan *precision* 0,96%.

## 4 Kesimpulan

Pada penelitian ini analisis sentimen mengenai opini survei Regsosek 2022 menggunakan 500 data *tweets* yang diambil dari tanggal 1 September 2022 sampai 30 November 2022. Metode *K-Nearest Neighbor* yang diterapkan untuk menganalisa sentimen masyarakat menghasilkan akurasi yang baik sebesar 96%, dengan *recall* dan *precision* sebesar 1,00 dan 0,96. Dimana, model dapat memprediksi 97 sentimen dengan tepat dari 101 data uji. Berdasarkan dari polaritas *tweets* yang didapat, opini masyarakat didominasi oleh opini negatif dengan kata yang sering muncul atau digunakan masyarakat untuk menyampaikan opininya adalah data Regsosek, tugas Regsosek dan sensus sosial ekonomi.

## Daftar Pustaka


[1] Badan Pusat Statistik Kota Lhokseunawe. *Pendataan Awal Registrasi Sosial Ekonomi (REGSOSEK) Tahun 2022*. Lhokseunawe: Badan Pusat Statistik Kota Lhokseunawe. 2022.

[2] Badan Pusat Statistik Kabupaten Kapuas. *Pendataan Awal Registrasi Sosial Ekonomi (REGSOSEK) Tahun 2022*. Kapuas: Badan Pusat Statistik Kabupaten Kapuas. 2022.

[3] Fauziyyah, A. K. Analisis Sentimen Pandemi Covid19 Pada Streaming *Twitter* Dengan Text Mining Python. *Jurnal Ilmiah SINUS,* 18(2):31, 2020.

[4] Rofiqoh, U., Setya, Perdana R., & Fauzi, M. A. Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada *Twitter* Dengan Metode Support Vector Machine dan Lexicon Based Features [Internet]. Vol. 1. 2017. Available from: http://j-ptiik.ub.ac.id

[5] Somantri, O. JEPIN (Jurnal Edukasi dan Penelitian Informatika) Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining. 2019; Available from: www.google.com/maps

[6] Furqan, M., & Mayang, S. S. Ilmu Komputer Fakultas Sains dan Teknologi P. Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia Sentiment Analysis using K-Nearest Neighbor towards the New Normal During the Covid-19 Period in Indonesia [Internet]. Vol. 21. 2022. Available from: www.tripadvisor.com

[7] Astar, N., Dewa G. H. D., & Gede, I. Analisis Sentimen Dokumen *Twitter* Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier. *Jurnal Sistem dan Informatika (JSI)*, 15(1):9-27, 2020.

[8] Mika, P. I. M., & Siahaan, D. Classification of Mobile Application Reviews using Word Embedding and Convolutional Neural Network. *Jurnal Ilmiah Teknologi Informasi*. *18*, 2019 .

[9] Gilyarovskaya, E. A. Automated classification of service reports using natural language processing techniques. 2021.

[10] Saadah, M. N., Atmagi, R. W., Rahayu D, S., & Arifin, A. Z. Sistem Temu Kembali Dokumen Teks dengan Pembobotan Tf-Idf Dan LCS. 2013.

[11] Hamzah, F., Astuti, W., & Purbolaksono, M. D. Sentiment Analysis pada movie review menggunakan Feature Selection Chi Square dan Support Vector Machine Classifier. *e-Proceeding of Engineering,* 9, 2022.

[12] Harun, R., Chandra P. K., & Lasena, Y. Penerapan Data Mining Untuk Menentukan Potensi Hujan Harian Dengan Menggunakan Algoritma K Nearest Neighbor (KNN) [Internet]. *Jurnal Manajemen informatika & Sistem Informasi*, 3, 2020. Available from: http://e-journal.stmiklombok.ac.id/index.php/misi

[13] Arora, I., Khanduja, N., & Bansal, M. *Effect of Distance Metric and Feature Scaling on KNN Algorithm while Classifying X-rays*. 2021.

[14] Kataria, A., Singh MD. International Journal of Emerging Technology and Advanced Engineering A Review of Data Classification Using K-Nearest Neighbour Algorithm [Internet]. Vol. 9001, Certified Journal. 2008. Available from: www.ijetae.com

[15] Yudhana, A., & Agus J. S. H. Algoritma K-NN Dengan Euclidean Distance Untuk Prediksi Hasil Penggergajian Kayu Sengon. *TRANSMISI*, 22(4), 2020. Available from: https://ejournal.undip.ac.id/index.php/transmisi

[16] Deviyanto, A., Didik, W. M. R., Informatika UIN Sunan Kalijaga Yogyakarta Jl Marsda Adi Sucipto No T. Penerapan Analisis Sentimen Pada Pengguna *Twitter* Menggunakan Metode K-Nearest Neighbor. *Jurnal Informatika Sunan Kalijaga*.

[17] Chicco, D., Tötsch, N., & Jurman, G. The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData*, 14:1–22, 2021.