Jurnal Ilmiah Wahana Pendidikan, Maret 2023, 9(5), 132-142
DOI: https://doi.org/10.5281/zenodo.7728573
p-ISSN: 2622-8327 e-ISSN: 2089-5364
Accredited by Directorate General of Strengthening for Research and Development
Available online at https://jurnal.peneliti.net/index.php/JIWP

Analisis Sentimen Komentar Video Youtube Dengan Metode K-Nearest Neighbor

Dickna Niken Larasakti[1], **Abdul Aziz**[2], **Danang Aditya**[3]

[1,2,3]Teknik Informatika, Universitas PGRI Kanjuruhan Malang

Received: *21 Januari 2023*
Revised: *28 Januari 2023*
Accepted: *8 Februari 2023*

Abstract

This study discusses the result of sentiment analysis of comments on Youtube videos using the K-Nearest Neighbor method. It contains various opinionsabout video content that have been watched. Opinnios given in comments can be used as an assessment and analyze how the sentiment rises. WhileYouTube only facilitates like and dislike buttons which can be seen from the number of clicks, the comments in a video will be used to analyze a sentiment that appears. Through this research, the system will classify each comment contained in the video, and make categories into positive and negative sentences. Before the classification results are obtained, it will go through several stages such as preprocessing, term weighting, similarity calculations and arrive at the calculation of accurate results. In addition to the calculation accuracy, there are also calculations of precision and recall. In this study using 4 scenarios with differences in the percentage of the amount of testing, training data and the value of K, with the aim of finding the best accuracy. The conclusion based on the results of the study is the large amount of training data, testing data and the value of K affects the accuracy results.The amount of testing data and training data also affects the accuracy search time. The highest accuracy is 92.71% with 3% testing data and 97% training data with k-7. The utilization of this KNN method has a good and accurate performance in the process of classifying YouTube video comments

Keywords: *Sentiment Analysis, YouTube, K-Nearest Neighbor, Accuracy, Python*

(*) Corresponding Author: *nikendickna@gmail.com[1)],abdul.aziz@unikama.ac.id[2)], anang.aditya@unikama.ac.id[3)]*

How to Cite: Larasakti, D., Aziz, A., & Aditya, D. (2023). Analisis Sentimen Komentar Video Youtube Dengan Metode K-Nearest Neighbor. *Jurnal Ilmiah Wahana Pendidikan, 9*(5), 132-142. https://doi.org/10.5281/zenodo.7728573

PENDAHULUAN

Perkembangan teknologi ini menghasilkan berbagai macam media sosial, salah satunya bernama Youtube. Selain mendapatkan informasi, penonton dapat memberikan *feedback* melalui kolom komentar yang mana berisi opini dan pendapat terhadap konten video yang dilihatnya. Opini yang diberikan dapat digunakan sebagai peniIaian dan menganaIisa bagaimana sentimen yang muncuI terhadap narasumber serta topik yang diperbincangkan. Dalam proses pengolahan data ini, cara pengelompokkan yang dapat digunakan bernama *text mining*. DaIam ugensi pengelompokkan opini yang sudah diuraikan, *sentiment analysis* menjadi *text mining* yang memiliki fungsi untuk mengelompokkan apakah suatu opini dapat dikatakan sebagai opini positif atau opini negatif.

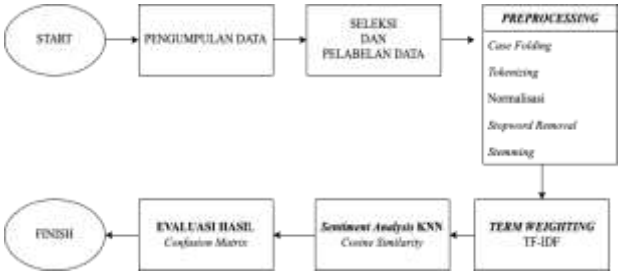
Proses klasifikasi opini ini diadopsi dalam bentuk analisis sentimen, oleh karena itu perlu adanya metode yang mampu mengklasifikasikan opini dengan tepat. Dalam penelitian ini metode yang digunakan adalah K-Nearest Neighbor (K-NN). K- Nearest Neighbor merupakan cara atau metode dimana suatu objek akan dikelompokkan berdasarkan data yang memiliki kemiripan atau jarak

terdekat dengan objek yang sudah ditetapkan. Dalam penerapannya nilai K digunakan untuk mencari tetangga paling dekat yang terlibat dalam proses menentukan prediksi label pada data uji. Setelah nilai terpiIih kemudian melakukan perhitungan kelas dari nilai K tetangga terdekat. Hasil keIas dengan jumIah perhitungan suara terbanyak yang akan dilakukan proses pemberian label sebagai label kelas hasil prediksi pada data uji tersebut (Reski, 2020).

Penggunaan metode KNN untuk klasifikasi teks menghasilkan akurasi yang tinggi dan berdasarkan penelitian terdahulu metode ini dapat dilakukan untuk mengklasifikasikan sebuah teks. Penelitian ini bertujuan untuk membuktikan bahwa dengan menggunakan data komentar ini dapat menghasilkan akurasi tinggi dengan menggunakan metode K- Nearest Neighbor. Data dalam penelitian ini menggunakan komentar video salah satu *channel* populer di Indonesia Deddy Corbuzier yang berisi *podcast* antara Deddy sebagai *host* dengan bintang tamunya sebagai narasumber. Merujuk kepada uraian yang telah dipaparkan, rumusan masalah yang diangkat adalah bagaimana model klasifikasi sentimen analisis komentar pada video youtube dan keakurasian metode K-Nearest Neighbor pada pengklasifikasian komentar youtube.

PEMBAHASAN

A. Rancangan Penelitian



Gambar 1 Rancangan Penelitian

B. Pengumpulan Data

Data yang digunakan adalah komentar yang didapat dari video youtube Deddy Corbuzier yang berjudul "INDONESIA DAN COVID19 (Menteri Sosial Juliari Batubara)". Pada proses ini, pengumpulan data komentar dilakukan dengan teknik *scrapping*, yaitu pengambilan data komentar youtube yang dilakukan dengan pemrograman bahasa Python.

Data yang diambil adalah data komentar dari bulan Juni sampai dengan bulan Oktober. Data diambil pada bulan Januari, pengambilan data ini berupa berupa nama pemilik akun *YouTube*, komentar dan waktu pemilik akun tersebut berkomentar. Data yang sudah didapat disimpan kedalam Microsoft Excel yang berjumlah 4.774 komentar.

Table 1 Hasil Pengumpulan Data

No	Author	Comment	Time
1.	Muhammad Ari	Nice,,, podcast berbobot dan asik ga kaku kaya kanebo kering. Padahal sama pak mentri. Pak mentrinya asik bro, om dedi oke yg di undang mentri2 atw.pejabat yg kaya gini	Juni
2.	Daffa Relics	Ok	Juni
3.	blitar kota	Jos	Juni
4.	Dataku Fakta	Cek channel gue bro, ada data menarik	Juni
5.	LutfilahDz	Asik, mantap om ded	Juni

C. Seleksi dan Pelabelan Data

Data diseleksi untuk menentukan data yang akan dipertahankaan untuk digunakan pada tahap selanjutnya serta penghapusan data yang mengalami duplikat dan yang tidak diperlukan. Data komentar yang sudah didapat kemudian diurai menjadi sebuah kalimat. Selanjutnya dilakukan pelabelan manual sesuai kategori yang sudah ditentukan, yaitu kalimat positif dan negatif.

Table 2 Hasil Seleksi dan Pelabelan

No	Author	Komentar	Kalimat	Label
1	Dataku Fakta	Cek channel gue bro, ada data menarik	Cek channel gue bro, ada data menarik	P
2	LutfilahDz	Asik, mantap om ded	Asik, mantap om ded	P

Pada proses ini dilakukan penghapusan kalimat yang *spamming* sehingga menghasilkan 4.774 komentar dan 8.122 kalimat yang sudah diuraikan. Selanjutnya dilakukan pelabelan secara manual sesuai kategori yang sudah ditentukan, yaitu kalimat positif dan negatif.

Table 3 Contoh data spamming yang dihapus

No	Author	Komentar	Kalimat
1.	SOMPYOH Channel	Next : Pak Jokowi	Next : Pak Jokowi
2.	Ilham channel	Next : Pak Jokowi	Next : Pak Jokowi
3.	Usaid Abdullah	mantap	mantap
4.	Nurul Huda	Mantap	Mantap
5.	Feri Angga	Up	Up

Dari 8.122 kalimat yang telah diambil, hanya menysahkan 6.490 kalimat yang akan dipakai untuk tahap berikutnya. Data yang akan digunakan pada klasifikasi sentimen untuk proses-proses selanjutnya hanyalah data kalimat saja.

Table 4 Data Kalimat Klasifikasi Sentimen

No.	**Kalimat**
1.	Nice,,, podcast berbobot dan asik ga kaku kaya kanebo kering
2.	Padahal sama pak mentri
3.	Pak mentrinya asik bro, om dedi oke yg di undang mentri2 atw
4.	pejabat yg kaya gini
5.	Cek channel gue bro, ada data menarik
6.	Asik, mantap om ded
7.	Terimakasih OM @DEDDYCORBUZIER

D. Preprocessing

Tahapan setelah peneliti mendapatkan data yang dibuuthkan yaitu menganalisis data yang sesuai dengan standar K-Nearest Neighbor.

a. *Case Folding* : Mengubah kata menjadi format yang sama yaitu diubah menjadi huruf kecil
b. *Tokenizing* : Sekumpulan Kalimat akan dipecah menjadi token dan menghilangkan tanda baca, pemisah kata, *emoji* dan karakter spesial.
c. Normalisasi : Menormalkan kalimat sehingga kalimat yang tidak baku menjadi kaIimat normal, sehingga bahasa yang tidak baku tersebut dapat dikenali sebagai bahasa yang sesuai dengan KBBI.
d. *Stopword Removal:* Menghapus kata yang tidak penting dan tidak ada pengaruhnya pada proses perolehan informasi
e. *Stemming :* Menghapus kata imbuhan awal dan akhir kalimat sehingga menjadi kata dasar.

E. *Term Weighting*

Dengan menggunakan TF/IDF, akan dilakukan pengukuran pembobotan suatu kata, yang akan menentukan klasifikasi uji data selanjutnya.Pada penelitian ini menggunakan *Tf-idf.* Kata yang digunakan adalah kata dari hasil *preprocessing* terakhir .

Table 5 Hasil Pembobotan TF

	TERM	**TF**								**df**	**Idf (*log (n/df)*)**
		D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**		
1	bagus	1								**1**	0,84509804
2	kaku	1								**1**	0,84509804

3	kaya	1		1						**2**	0,54406804
4	kering	1								**1**	0,84509804

Pada tabel menghitung jumlah kata yang muncul pada setiap dokumen, sehingga dalam perhitungan nya dapat memberi nilai satu atau jika lebih maka dapat ditambahkan sesuai kondisi kemunculan suatu kata. Kata yang digunakan adalah kumpulan dari data yang sudah melalui proses *preprocessing.* Sebelum mencari TF-IDF, perlu mencari nilai TF, DF dan IDF.

Setelah nilai TF dan DF diketahui,selanjutnya menghitung IDF. Setelah didapati nilai IDF maka selanjutnya memasukkan nilai TF-IDF berdasarkan kelompoknya

Table 6 Hasil Pembobotan Kata TFIDF

		TF							
	TERM	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**
1	bagus	0,84509804							
2	kaku	0,84509804							
3	kaya	0,54406804		0,54406804					
4	kering	0,84509804							

F. *Cosine Similarity*

Memiliki kegunaan untuk menguji kedekatan antara dokuemn dengan contoh uji coba. Selain itu *relevansi query* dan kedekatannya diukur berdasrkan vektor apakah memiiki kedekatan atau tidak.

$$CosSim\ (x, dj) = \frac{\sum_{i=1}^{m} xi.\ dji}{\sqrt{\sum_{i=1}^{m} x^2 i}\ .\ \sqrt{\sum_{i=1\ d^2}^{m} ji}} \qquad (4)$$

Keterangan:

X : uji dokumen

dj : sampel dokumen

xi : term i dalam bobot pada uji dokumen

dji : term i dari bobot terhadap sampel dokumen

G. *Confussion Matrix*

Proses mengukur parameter ini digunakan untuk pengevaluasi hasil dari sebuah metode yang terlah dilakukan proses klasifikasi. Dalam penelitian ini proses evaluasi menggunakan *Confussion Matrix*.

Table 7 Confussion Matrix

		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

True Positive (TP) : Jumlah prediksi benar dan fakta benar
True Negative (TN) : Jumlah prediksi salah dan fakta salah
False Positive (FP) : Jumlah prediksi benar dan fakta salah
False Negative (FN) : Jumlah prediksi salah dan fakta benar

$$accuration \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

$$precission \frac{TP}{TP+FP} \quad (6)$$

$$recall \frac{TP+TN}{TP+FN} \quad (7)$$

Dari hasil matrix tersebut dapat disimpulkan terdapat performasi dari proses klasifikasi berupa accuracy. Accurcy sendiri adalah ratio dari jumlah ketepatan prediksi yang ada dalam K.

H. Klasifikasi KKN

Gambar 2 Proses Klasifikasi KKN

Identifikasi data latih dan data uji pada data komentar dilakukan secara acak sebanyak empat kali percobaan dengan persentase yang berbeda, yaitu (3% dan 97%), (10% dan 90%), (27% dan 73%), (35% dan 65%). Pada tahap ini data yang digunakan adalah data yang sudah dipecah menjadi data latih dan data uji yang telah diberi label positif dan negatif. Metode ini menggunakan *cosine similarity* untuk mengukur jarak dokumen. Nilai *similarity* diperoleh dengan membagi bobot dikalikan dengan perkalian vektor data uji dengan vektor data latih.

Tahap selanjutnya hasil dikelompokkan berdasarkan kategori dan diurutkan dalam urutan menurun, nilai yang lebih tinggi menunjukkan bahwa tes dan pelatihan serupa. Setelah mengurutkan nilai *Simmilarity* dari terbesar ke terkecil, kemudian menentukan kelompok data uji berdasarkan nilai k. Hasil pengujian telah dilakukan untuk mendapat K yang optimal yang menghasilkan akurasi. Berdasarkan uji coba pembagian data *training* dan data *testing* dapat menghasilkan nilai k dari masing –masing presentase. Maka nilai K dari presentase data *training* dan data *testing* memiliki nilai k optimal yang berbeda. Setiap skenario memiliki jumlah data *testing* dan data *training* yang berbeda dan memiliki nilai k yang dieksekusi sama sebanyak 2 kali.

Table 8 Skenario Pengujuan

Data		**K**
Training	Testing	
3%	97%	5
10%	90%	5
27%	73%	5
35%	65%	5
37%	63%	5
3%	97%	7
10%	90%	7
27%	73%	7
35%	65%	7
37%	63%	7

I. Hasil Akurasi KKN

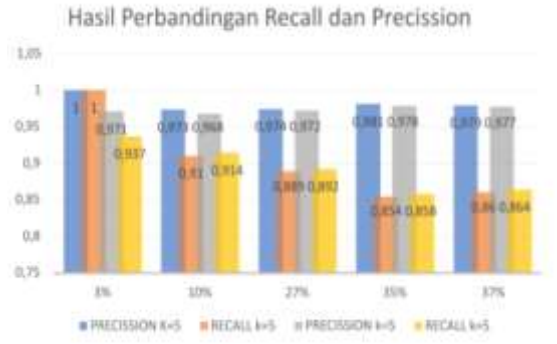
Penerapan metode K Nearest Neighbor pada penelitian ini digunakan untuk mengklasifikasikan komentar pada video youtube sebagai sentimen positif dan sentimen negatif. *Confussion matrix* digunakan untuk pengujian akurasi pada penelitian ini. . Untuk nilai k yang diuji dengan masing masing presentase adalah k=5 dan k=7.

Table 9 Hasil Pengujian KKN

Testing	HASIL			
	AKURASI	**K**	**PRECISSION**	**RECALL**
3%	92,715	7	1	1
10%	89,13	7	0,9735	0,9107
27%	87,132	7	0,974	0,889
35%	84,24	7	0,981	0,854
37%	84,648	7	0,979	0,86
3%	91,39	5	0,971	0,937
10%	88,933	5	0,968	0,914
27%	87,279	5	0,9724	0,8926
35%	84,52	5	0,9786	0,8586
37%	85,02	5	0,9779	0,8643

Gambar 3 Grafik Perbandingan Nilai Akurasi

Seperti terlihat pada gambar di atas, hasil pengujian pada masing-masing data *training* dan data *testing* dengan representasi yang berbeda memberikan akurasi yang berbeda. Begitu pula dengan perbedaan nilai K, memiliki hasil yang berbeda namun tidak signifikan. Nilai akurasi tertinggi dengan nilai 92,715% dengan data *testing* 3% dan data *training* 97% dengan nilai k=5. Berdasarkan hasil grafik tersebut, semakin sedikit data *testing* hasil akurasi nya semakin tinggi. Namun, untuk perbedaan nilai K, tidak ada perbedaan yang signifikan, nilai K=5 menghasilkan akurasi yang lebih tinggi.



Gambar 4 Grafik perbandingan recall dan precission

Setelah perbandingan hasil dengan nilai *accuration, precission,* dan *recall*, penulis juga membandingkan waktu dalam proses jalannya program dalam pencarian nilai *accuration.* Berikut ditampilkan waktu dalam proses perhitungan *accuration.*

Table 10 Perbandingan waktu pencarian akurasi

Testing	AKURASI k=5	Time (s) k=5	AKURASI k=7	Time (s) k=7
3%	92,715	3	92,715	3
10%	89,13	7	89,13	7
27%	87,132	16	87,132	16
35%	84,24	19	84,24	19
37%	84,648	19	84,648	19

Gambar 5 Grafik Perbandingan Waktu Akurasi

Dari gambar grafik diatas, dapat kita lihat adanya perbedaan waktu dalam pencarian akurasi. semakin banyak data *testing* membutuhkan waktu yang lama. Waktu tercepat dicapai oleh data *testing* 3% dan data *training* 97% dengan waktu 3detik. Namun dari perbedaan nilai K, tidak ada perubahan, yaitu k=5 dan k=7 memiliki waktu yang sama dalam proses pencarian akurasi. Berdasarkan gambar grafik, semakin sedikit data *testing*, waktu yang dibutuhkan untuk menghitung akurasi semakin sedikit pula. Namun, untuk perbedaan nilai K tidak ada perbedaan waktunya

KESIMPULAN

Pada proses klasifikasi KNN data dibagi menjadi data *testing* dan data *training* dengan presentase yang berbeda yaitu (3% dan 97%), (10% dan 90%), (27% dan 73%), (35% dan 65%), (37% dan 63%). Untuk nilai k yang diuji dengan masing masing presentase adalah k=5 dan k=7.

Berdasarkan hasil pengujian ini , dapat disimpulkan sebagai berikut :

1. Hasil pengujian pada masing-masing data *training* dan data *testing* dengan representasi yang berbeda memberikan akurasi yang berbeda. Demikian pula dengan perbedaan nilai K, memiliki hasil yang berbeda namun tidak signifikan.

2. Semakin sedikit data *testing*, waktu yang dibutuhkan untuk menghitung akurasi semakin cepat. Namun untuk perbedaan nilai K tidak ada perbedaan waktu.
3. Semakin banyak data *training* , hasil akurasi yang dihasilkan semakin tinggi. Hasil akurasi tertinggi yaitu 92,71% dengan 3% data *testing* dan 97% data *training* dengan k=7.
4. Penggunaan metode K-Nearest Neighbor memiliki kinerja yang baik dan akurat dalam pengklasifikasian analisis sentimen komentar pada video youtube.

Adapun saran yang dapat membantu penelitian untuk lebih baik dan berkembang :

1. Menggunakan data lebih banyak, karena semakin banyak data yang menghasilkan akurasi yang lebih baik.
2. Menggunakan lebih banyak atribut untuk hasil yang lebih baik dan berbeda dengan tingkat akurasi yang lebih tinggi.
3. Dimungkinkan untuk berksperimen dengan metode lain yang dapat membantu algoritma K-NN mencapai hasil yang lebih baik dan akurat.

DAFTAR PUSTAKA

Akhmad Deviyanto, M. Didik R. Wahyud. 2018. "PENERAPAN ANALISIS SENTIMEN PADA PENGGUNA TWITTER MENGGUNAKAN METODE K-NEAREST

Candra, Reski Mai Rozana, Anindya Nanda. 2020. "Klasifikasi Komentar Bullying pada Instagram Menggunakan Metode K-Nearest Neighbor." *IT Journal Research and Development (ITJRD)* 45-52.

Efendi, Zuliar Mustakim. 2017. "Text Mining Classification Sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi." *Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 9* 18-19.

Feldman, Ronen Sanger, James. 2006. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.*

Fitrianti, Risma Putri Kurniawati, Ana Agusten, Dina. 2019. "Implementasi Algoritma K - Nearest Neighbor Terhadap Analisis Sentimen Review Restoran Dengan Teks Bahasa Indonesia Risma." *SNATi* 27-32.

Muslimah, Nurul Indriati Wihandika, R.C. 2019. "Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbor (K-NN)." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 196-204.

Riany, Jane Fajar, Mohammad Lukman, Musfirah Putri. 2016. "Penerapan Deep Sentiment Analysis pada Angket Penilaian Terbuka Menggunakan K-Nearest Neighbor." Jurnal Sisfo Vol. 06 No. 01 147-156.

Salam, Abu Zeniarja, Junta Khasanah, Rima Septiyan Uswatun. 2018. "Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa Ekspedisi Barang J&T Ekpress Indonesia)." *Prosiding SINTAK* 480- 486.

Siregar, Zuhdiyyah Ulfah, Riki Ruli Siregar, dan Rakhmat Arianto. 2019. "Klasifikasi Sentiment Analysis Pada Komentar Pseta Diklat Menggunakan Metode K-Nearest Neighbor." *Jurnal Kilat.*

Suprapto, Aji. 2017. "SISTEM KLASIFIKASI OPINI PENGGUNA MASKAPAI PENERBANGAN DI INDONESIA PADA JEJARING SOSIAL TWITTER MENGGUNAKAN METODE K-NEAREST NEIGHBOR."

Windiarti, N. R. 2018. "Klasifikasi Opini Netizen Berbahasa Indonesia Berbasis Twitter Menggunakan Metode Improved K-Nearest neighbor".