Bit (Fakultas Teknologi Informasi Universitas Budi Luhur) p-ISSN: 1693-9166
Vol. 19, No. 2, September 2022, hlm. 98-103 e-ISSN: 2685-127X

ANALISIS SENTIMEN KESEHATAN MENTAL MENGGUNAKAN *K-NEAREST NEIGHBORS* PADA SOSIAL MEDIA *TWITTER*

Mahesworo Langgeng Wicaksono[1*], Rusdah[2], Diwi Apriana[3]

[1]Fakultas Teknologi Informasi, Sistem Informasi, Universitas Budi Luhur, Jakarta, Indonesia
[2]Fakultas Teknologi Informasi, Magister Ilmu Komputer, Universitas Budi Luhur, Jakarta, Indonesia
[3]Fakultas Teknik, Ilmu Komputer, Universitas Pat Petulai, Bengkulu, Indonesia
Email: [1]wicaksonomahes@gmail.com, [2]rusdah@budiluhur.ac.id, [3]dapriana102@gmail.com

(Naskah masuk: 12 Agustus 2022, diterima untuk diterbitkan: 29 Agustus 2022)

Abstrak

Isu kesehatan mental masih menjadi satu permasalahan kesehatan yang signifikan di dunia modern. Pemahaman dan stigma yang kurang baik serta kesadaran kesehatan mental yang rendah turut andil dalam upaya penyuluhan perihal kesehatan mental. Isu tentang kesehatan mental banyak dibahas pada media sosial, salah satunya Twitter. Sehingga perlu dilakukan analisis sentimen terhadap isu kesehatan mental pada Twitter. Dataset pada penelitian ini menggunakan ulasan masyarakat pada tanggal 16 Mei 2022 dengan kata "*Mental Health"*. Metode penelitian yang digunakan pada penelitian ini ada beberapa tahap, seperti melakukan pengolahan data menggunakan algoritma *K-Nearest Neighbors* dengan melakukan perbandingan algoritma klasifikasi *Support Vector Machine* dan *Decision Tree*. Proses pengolahan data penelitian menggunakan tools *Rapid Miner*. Kesimpulan penelitian ini adalah, berdasarkan hasil eksperimen dengan dataset ulasan sentimen positif sebanyak 639 dan ulasan sentimen negatif sebanyak 193, maka hasil pemrosesan modeling dengan menggunakan algoritma *K-Nearest Neighbors* didapatkan hasil terbaik saat menggunakan metode *split data* 70:30 dengan nilai k berada pada angka 5, yaitu menghasilkan precision 60.87%, recall 44.03% dan accuracy 58.39%.

Kata kunci: *analisis sentimen, text mining, k-nearest neighbors, kesehatan mental*

SENTIMENT ANALYSIS OF MENTAL HEALTH USING K-NEAREST NEIGHBORS ON SOCIAL MEDIA TWITTER

Abstract

Mental health issues are still significant health problems in the modern world. Poor understanding, stigma, and low mental health awareness contribute to efforts to educate people about mental health. The issue of mental health is widely discussed on social media, one of which is Twitter. So it is necessary to analyze sentiment on mental health issues on Twitter. The dataset in this study uses community reviews on May 16, 2022, with the search word "Mental Health." The research method used in this study has several stages, such as processing data using the K-Nearest Neighbors algorithm by comparing the Support Vector Machine classification algorithm and Decision Tree Processing research data using Rapid Miner tools. The conclusion of this study is, based on experimental results with a dataset of 639 positive and 193 negative sentiment reviews. The results of modeling processing using the K-Nearest Neighbors algorithm obtained the best results when using the split data method 70:30 with k value at number 5, producing precision of 60.87% and recall of 44.03%, respectively, and accuracy of 58.39%.

Keywords*: sentiment analysis, text mining, k-nearest neighbors, mental health*

1. PENDAHULUAN

Kesehatan mental adalah sebuah kondisi di mana Anda bisa tenang, menikmati kehidupan sehari-hari, dan berterima kasih kepada orang lain. Orang dengan kesehatan mental dapat memaksimalkan potensinya untuk menghadapi tantangan hidup. Hubungan positif dengan orang lain [1]. Kesehatan mental atau yang sering disebut dengan kesehatan jiwa merupakan bagian yang tidak terpisahkan yang juga merupakan keadaan individu merasa damai dan sehat secara *mental*, emosional dan sosial yang mempengaruhi cara berpikir, perasaan, perilaku, pengambilan keputusan, mengatasi stres, dan interaksi sosial dengan orang lain[2]. Berbagai upaya rehabilitasi dan diskusi akademis telah dilakukan oleh organisasi terkait untuk mengurangi prevalensi penyakit mental Ini tidak berhasil ketika masyarakat berkontribusi

pada diskriminasi terhadap mereka yang terkena dampak, dan pemahaman dan stigma yang rendah tentang kesehatan mental, dan kesadaran yang rendah berkontribusi pada upaya penyembuhan kesehatan mental, Secara umum stigma penyakit jiwa dapat dikonseptualisasikan sebagai stigma publik dan stigma diri. [3]

Berkaitan dengan isu kesehatan *mental* ialah banyak dari para pengguna *twitter* yang dikalangan generasi *milenial* dan *Gen-Z* yang tingkat kepeduliannya mereka menampilkan *headline* di banyak media sosial. Data yang didapat dari *Platform Twitter* pada tahun 2021 menampilkan bahwa ada kenaikan 17% tingkat percakapan dan pembahasan perihal kesehatan mental dalam rentang tahun 2018 hingga 2021, *Platform Twitter* juga berkomitmen melakukan banyak ide solusi dalam menangani perihal isu ini, salah satunya ialah bermitra dengan otoritas dan organisasi *non-profit* kesehatan mental di Asia Tenggara untuk memahami percakapan publik mengenai kesehatan mental, melakukan advokasi dan *campaign*, serta meluncurkan *#ThereIsHelp* layanan notifikasi yang menyediakan sumber daya dan informasi berharga mengenai kesehatan mental dan kita bisa mengakses lebih lanjut perihal kesehatan *mental* yang diekspresikan oleh para pengguna didalam sebuah *platform twitter* [4].

Pada penelitian ini akan dilakukan pengklasifikasian dari data *tweet* yang didapat tingkat akurasinya dengan menggunakan algoritma *K-Nearest Neighbors* yang sudah diperbandingkan dengan algoritma *Decision Tree* dan *Support Vector Machine.* Algoritma *K-Nearest Neighbors* adalah contoh dasar non-konstruktif, pengklasifikasi berdasarkan representasi deklaratif eksplisit dari suatu kategori, tetapi seperti pengujian, kategori yang dilampirkan pada dokumen pelatihan materi yang bergantung pada label merupakan metode untuk mengklasifikasikan objek berdasarkan data latih yang mewakili jarak terpendek [5].

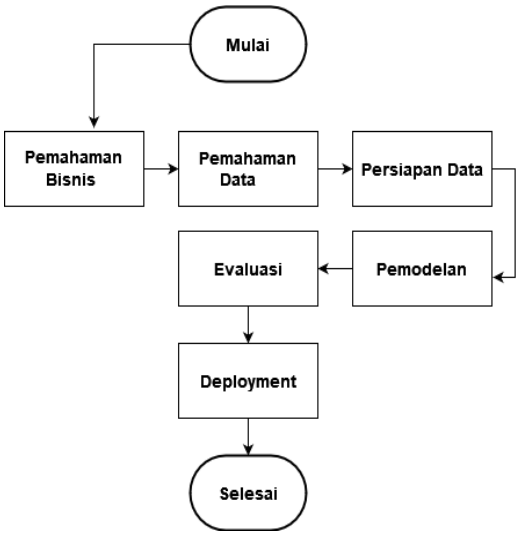
Penelitian sebelumnya yang menggunakan data dari Twitter terkait kesehatan mental, sudah dilakukan dengan menggunakan algoritma *Support Vector Machine* [6] yang didapatkan dengan *accuracy* tertinggi sebesar 80.81%. Penelitian lainnya dengan algoritma *Naïve Bayes* [7] mendapatkan akurasi *naïve bayes* sebesar 79%. dan penelitian [8] mendapatkan tingkat *Accuracy Naïve Bayes* sebesar 89%.

Dari penjabaran diatas, peneliti menganalisis mengenai isu kesehatan mental pada data tweet yang telah didapatkan, kemudian diolah dan diklasifikasikan dengan metode *K-Nearest Neighbors*. Adapun tujuan dari penelitian ini mengetahui stigma yang berkembang dalam masyarakat mengenai kesehatan mental dari media sosial twitter, dan untuk mengetahui tingkat akurasi yang dihasilkan dari algoritma *K-Nearest Neighbors* dalam mengklasifikasikan sentimen yang berisi tanggapan masyarakat di media sosial twitter yang berkaitan dengan isu kesehatan mental, serta hasil penelitian ini dapat menambah pengetahuan dalam menganalisis komentar pada media sosial twitter dan dapat dijadikan sebagai bahan referensi untuk evaluasi dan acuan dalam peningkatan kebijakan lembaga kesehatan mental.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini terdiri dari beberapa tahapan yang ditampilkan pada gambar 1. Penelitian ini menerapkan metodologi *Cross-Industry Standard Process For Data Mining* (CRISP-DM) [9]



Gambar 1. Tahapan Penelitian Analisis Sentimen Kesehatan Mental Menggunakan K-Nearest Neighbors Pada Sosial Media Twitter

2.2 Tahapan CRISP-DM

a. *Business Understanding*

Pada tahapan ini, dilakukan pemahaman mengenai permasalahan yang ingin diangkat yaitu analisis sentimen terhadap isu kesehatan mental pada media sosial Twitter.

b. *Data Understanding*

Dataset penelitian bersumber dari Twitter, yaitu ulasan masyarakat pada tanggal 16 Mei 2022 dengan kata kunci “mental health” dan “Kesehatan Mental”. Jumlah dataset 5.000 tweet. Proses pengambilan dataset penelitian menggunakan software Rapid Miner. Kemudian dataset diolah untuk memilih atribut yang ingin digunakan pada penelitian dengan cara melakukan *select attribute* pada Rapid Miner. Tahap berikutnya adalah *remove duplicates* menggunakan Microsoft Excel untuk menghapus data yang muncul secara berulang. Kemudian Dataset penelitian kemudian diberikan kepada pakar bahasa bernama Hanafi Mazi Syahputra S.Pd untuk memberikan validasi apakah ulasan pada dataset mengandung sentimen positif atau sentimen negatif. Tabel 1 adalah contoh dari hasil pemberian label sentimen positif dan negatif.

Tabel 1 Contoh *Labeling* Sentimen

Text	Sentimen
Tahun lalu ngobrol sm dua psikolog nemeni orang gitu trs ada obrolan kita kalo *mental health* ga bisa dianggap enteng "dipermainkan" tapi jangan sampai *mental health* orang lain mempermainkan kamu dgn dalih kamu mau dia *feel better* saat kamu malah ga ngerasa *its fine with urself*	*Positive*
Buat apa gue pulang ke rumah kalau cuman buat merusak *mental health* gue ahahahaha	*Negative*

c. *Data Preparation*

Tahapan selanjutnya adalah *preprocessing*. Proses ini meliputi beberapa tahapan, yang meliputi *transform cases, tokenize, filter tokens, filter stopwords dan stem*. Data yang telah dikumpulkan, selanjutnya dilakukan proses *preprocessing* menggunakan *Ms. Excel* dan juga bantuan *tool Rapid Miner*. Berikut tahapan yang terdapat dalam *data preparation*:

1. *Transform cases*, proses mengubah semua huruf pada data komentar menjadi huruf kecil. Setelah itu dilakukan kembali proses *remove duplicate* dan menghasilkan 2.302 komentar yang terdiri dari 1.886 komentar positif atau sebesar 82% dan 416 komentar negatif atau sebesar 18%.
2. *Tokenize*, proses memecah kalimat pada data komentar agar sistem dapat melakukan pengecekan satu persatu pada tiap kata yang terdapat dalam kalimat. Dalam proses ini digunakan operator *tokenize* dengan memilih *mode non letters*. Hal ini dilakukan agar token yang terbentuk nantinya hanya yang mengandung huruf saja. Pada proses ini menghasilkan sebanyak 4.841 atribut kata.
3. *Filter tokens*, proses untuk menyaring hasil token berdasarkan panjang karakter atau jumlah minimal huruf yang terdapat dalam satu kata. Pada proses ini menggunakan *operator filter tokens (by length)* dan mengubah *minimal chars* menjadi 3. Dari proses ini menghasilkan sebanyak 4.678 atribut kata.
4. *Stemming*, proses untuk menghilangkan imbuhan yang terdapat pada sebuah kata dan merubahnya menjadi bentuk kata dasar. Selain menghilangkan imbuhan, dilakukan pula mengubah *slang words* menjadi kata baku dan menyamaratakan kata-kata yang sama dengan penulisan yang bervariasi. Dalam mengubah token menjadi kata dasar dan mengubah beberapa *slang words*, peneliti membuat kamus secara manual dengan melihat Kamus Besar Bahasa Indonesia (KBBI) yang diakses secara *online*. Hal ini dilakukan karena *Rapid Miner* tidak menyediakan kamus kata dasar berbahasa Indonesia. Dalam proses ini menghasilkan sebanyak 2.971 atribut kata.
5. *Filter Stopwords*, proses untuk mengambil kata-kata penting dari hasil *tokenizing* dan membuang kata yang dapat diabaikan, atau menghilangkan kata yang tidak memiliki makna. Sebanyak 2.334 kata yang terfilter.

d. *Modeling*

Pada tahapan ini peneliti akan melakukan pemodelan terhadap dataset yang sudah dilakukan *preprocessing*. *Dataset* tersebut akan dipecah menjadi data *training* dan data *testing*, pada tahapan ini akan dipecah menjadi dua tahapan yaitu *split data* dan *cross fold validation*. Pada *split data* peneliti akan membagi *dataset* tersebut dengan perbandingan data *training* dan data *testing* yaitu 60:40, 70:30, dan 80:20. Lalu pada tahap *Cross Validation* dengan pembagian secara acak ke dalam 10 bagian (*number of folds* =10) *Cross Validation* adalah teknik untuk mengevaluasi atau memverifikasi tingkat keakuratan Model yang dibuat berdasarkan kumpulan *data* tertentu. Pemodelan tipikal Tujuannya adalah untuk memprediksi dan mengklasifikasikan *data* baru yang mungkin[10]. Pada tahap pemodelan ini peneliti menggunakan algoritma *K-Nearest Neighbors* namun sebelumnya telah dilakukan perbandingan model klasifikasi dengan menggunakan algoritma *Decision Tree* dan *Support Vector Machine*.

e. *Evaluation*

Pada tahapan ini peneliti akan melakukan evaluasi metode klasifikasi dengan mengukur performa menggunakan *confusion matrix* terhadap algoritma *K-Nearest Neighbors*.

f. *Deployment*

Pada tahapan ini adalah memberikan kesimpulan terhadap hasil yang didapat dan meninjau dari teori dengan permasalahan yang dihadapi

3. HASIL DAN PEMBAHASAN

3.1 *Data Preparation*

Setelah dilakukan tahap pengumpulan data, selanjutnya adalah data preparation di dalamnya merupakan tahap data preprocessing yang terdiri dari tahapan yang meliputi *transform cases, tokenize, filter tokens, filter stopwords dan stem*.(Tabel 2)

3.2 Modeling

Proses modeling pada penelitian ini menggunakan algoritma klasifikasi *K-Nearest Neighbors*. Namun, penelitian ini juga melakukan perbandingan dengan algoritma *Support Vector Machine* dan *Decision Tree*. Proses *modeling* dilakukan menggunakan metode Split Data, dan *K-Fold Cross Validation*

dengan 10-*fold cross validation*, yaitu membagi data keseluruhan menjadi 10 bagian.

1. Modeling *K-Nearest Neighbors* dengan Metode Split Data.

Berdasarkan Tabel 3, hasil perbandingan ketiga algoritma menggunakan metode split data, dengan perbandingan data testing dan data validasi sebagai berikut: 60:40, 70:30, dan 80:20. Hasil eksperimen menunjukkan bahwa algoritma KNN lebih unggul dibandingkan kedua algoritma lainnya. Kinerja algoritma KNN unggul pada pengujian dengan perbandingan 70:30 yaitu sebesar 58.39%.

Tabel 2. Tahapan *Preprocessing*

Preprocessing	**Sebelum**	**Sesudah**
Transform Cases	Lalu Nyatanya keluarga gw ga peduli sm *mental health*	lalu nyatanya keluarga gw ga peduli sm *mental health*
Tokenize	lalu nyatanya keluarga gw ga peduli sm *mental health*	“lalu” “nyata” “keluarga” “gw” “ga” “cukup” “peduli” “sm” “mental” “*health*”
Filter Tokens (by Length)	“lalu” “nyata” “keluarga” “gw” “ga” “cukup” “peduli” “sm” “mental” “health”	“lalu” “nyata” “keluarga” “cukup” “peduli” “mental” “health”
Filter Stopwords (Dictionary)	“lalu” “nyata” “keluarga” “cukup” “peduli” “mental” “health”	“nyata” “keluarga” “peduli” “mental” “health”
Stem (Dictionary)	“lalu” “nyata” “keluarga” “cukup” “peduli” “mental” “health”	“lalu” “nyata” “keluarga” “cukup” “peduli” “mental” “health”

Tabel 3. Perbandingan Hasil Akurasi dengan Pemodelan Split Data

Rasio Perbandingan	**Algoritma**		
	K-Nearest Neighbors	***Decision Tree***	***Support Vector Machine***
60:40	55.01%	51.05%	56.41%
70:30	**58.39%**	49.69%	56.83%
80:20	53.27%	52.34%	57.94%

Berdasarkan Tabel Pengujian pada algoritma K-Nearest Neighbors dilakukan sebanyak 5 kali dengan pembagian data 70:30. Pengujian dilakukan berdasarkan nilai k, yaitu k=1,k=3,k=5,k=7,k=9 pada Tabel 4 ini menggunakan angka k ganjil dikarenakan algoritma K-Nearest Neighbors bekerja dengan cara menentukan kelas berdasarkan kelompok mayoritas hasil dari pemilihan tetangga terdekat sebanyak k tetangga menjadi penentu kelas dari data uji. Nilai k=5 menghasilkan *accuracy* lebih besar dari pada k=1, k=3, k=7 dan k=9 dengan tingkat akurasi 58.39%. karena jumlah sentimen yang

kemunculannya paling banyak pada k=5 terdapat nilai perhitungan terbesar.

Tabel 4. Perbandingan dengan Nilai K

Nilai K	Akurasi
1	51.86%
3	55.90%
5	58.39%
7	56.21%
9	54.97%

2. Modeling dengan *cross validation*

Pada tabel 5 merupakan penjabaran perbandingan ketiga algoritma untuk mengukur akurasi dengan menggunakan metode *Cross Validation* yang memakai 10 *number of folds*. Hasil akurasi menggunakan *cross validation* menunjukan bahwa *K-Nearest Neighbors* lebih unggul dibanding algoritma *Decision Tree* dan *Support Vector Machine.*

Tabel 5. Perbandingan Hasil Akurasi dengan *Cross Validation*

Algoritma	Akurasi
Support Vector Machine	56.03%
K-NN	57.24%
Decision Tree	51.82

Tabel 6 menampilkan hasil proses pengujian algoritma *K-Nearest Neighbors* menggunakan *confusion matrix* dengan dengan *dataset* yang berjumlah 529 positif dan 542 negatif. Menggunakan model *split data* 70:30 ketika nilai k berada pada angka 5. Proses Pengujian menggunakan 749 *data training* dan 332 *data testing*:

Tabel 6. *Confussion Matrix*

	True Positive	***True Negative***	***Class Precision***
Pred Positive	45	70	60.87%
Pred Negative	118	89	57.00%
Class Recall	44.03%	72.39%	

Precision = $\frac{TP}{TP+TN} = \frac{45}{45+89}$ = 60.87%

Recall = $\frac{TP}{TP+FN} = \frac{45}{45+118}$ = 44.03%

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{45+89}{45+89+70+118}$ = 58.39%

Precision sebesar 60.87% merupakan persentase sentimen yang benar positif dari data keseluruhan komentar yang diprediksi positif. Sedangkan *precision* sebesar 57.00% adalah persentase sentimen yang benar negatif dari keseluruhan komentar yang diprediksi negatif. *Recall* sebesar 44.03% merupakan persentase sentimen yang diprediksi positif dibandingkan dengan keseluruhan komentar yang sebenarnya positif. Sedangkan *recall* 72.39% merupakan sentimen yang diprediksi negatif dibandingkan dengan keseluruhan komentar yang sebenarnya negatif.

4. KESIMPULAN

Berdasarkan hasil yang penelitian yang didapat, maka dapat diambil kesimpulan bahwa penelitian tentang analisis sentimen publik pada media sosial *twitter* terhadap isu *mental health* yang mengambil data dari media sosial *twitter* pada tanggal 16 Mei 2022 dengan jumlah dataset yang digunakan 1071 data dengan sentimen positif 529 data dan sentimen negatif 542 data berdasarkan notasi dari pakar, didominasi oleh sentimen negatif dan proses analisis sentimen dilakukan dengan menggunakan perbandingan terhadap metode klasifikasi *K-Nearest Neighbors*, *Support Vector Machine* dan *Decision Tree* yang menghasilkan akurasi nilai tertinggi yaitu dengan menggunakan algoritma *K-Nearest Neighbors*. Algoritma *K-Nearest Neighbors* mendapatkan akurasi terbesar dengan menggunakan *split data* 70:30 yang didapat ketika nilai k berada pada angka 5. Menghasilkan *precision* sebesar 60.87% *recall* sebesar 44.03% dan *accuracy* sebesar 58.39%.

DAFTAR PUSTAKA

[1] Direktorat Promosi Kesehatan dan Pemberdayaan Masyarakat Kementerian Kesehatan Indonesia, "Pengertian Kesehatan Mental," *promkes.kemkes.go.id*, 2018. https://promkes.kemkes.go.id/pengertian-kesehatan-mental (accessed Sep. 08, 2022).

[2] M. S. Muli, *Perancangan Media Kampanye Sosial Mental Health Berbasis Video Motion Comic Sebagai Upaya Menjaga Kejiwaan Para Remaja Pasca Pandemi Covid-19*, No. 8.5.2017. Universitas Dinamika, 2022.

[3] K. Aulia and L. Amelia, "Analisis Sentimen Twitter Pada Isu Mental Health Dengan Algoritma Klasifikasi Naive Bayes," *Siliwangi J. (Seri Sains Teknol.*, vol. 6, no. 2, pp. 60–65, 2020.

[4] Annisa Dewi Lestari, "Twitter: Obrolan soal Kesehatan Mental Naik Signifikan selama Pandemik," *idntimes.com*, 2021. https://www.idntimes.com/news/indonesia/annisa-dewi-lestari/twitter-obrolan-soal-kesehatan-mental-naik-signifikan-selama-pandemik?page=all (accessed Oct. 08, 2022).

[5] A. D. Adhi Putra, "Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.

[6] D. A. Pangestu, "Analisis Sentimen Terhadap Opini Publik Tentang Kesehatan Mental Selama Pandemi Covid-19 Di Media Sosial Twitter Menggunakan Naive Bayes Classifier Dan Support Vector Machine," *Jur. Stat. Fak. Mat. Dan Ilmu Pengetah. Alam Univ. Islam Indones. Yogyakarta*, 2020.

[7] K. Yan, D. Arisandi, P. Studi, S. Informasi, and U. Tarumanagara, "Analisis Sentimen Komentar Netizen Twitter Terhadap Kesehatan Mental Masyarakat Indonesia," *Junral Ilmu Komput. dan Sist. Inf.*, vol. 10, no. 1, pp. 1–8, 2022, [Online]. Available: https://journal.untar.ac.id/index.php/jiksi/article/view/17865.

[8] R. A. Yunis Femilia Nugraini, Rd. Rohmat Saedudin, "Implementasi Data Mining Dalam Kasus Mental Health Pada Sosial Media Twitter Menggunakan Metode Naive Bayes," vol. 8, no. 5, pp. 9260–9265, 2021.

[9] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

[10] A. Imron, "Analisis Sentimen Terhadap Tempat Wisata di Kabupaten Rembang Menggunakan Metode Naive Bayes Classifier," *Tek. Inform.*, pp. 10–13, 2019, [Online]. Available: https://dspace.uii.ac.id/handle/123456789/14268.