

METODE K- NEAREST NEIGHBOR UNTUK ANALISIS SENTIMEN TENTANG PENGHAPUSAN UJIAN NASIONAL

Faza Ahluna¹, Chelvin Joines Tutuarima², Imam Santoso³

¹Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, ²Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika,

³Universitas Teknologi Muhammadiyah Jakarta

¹faza@stikomcki.ac.id, ²chelvin@stikomcki.ac.id, ³imam.santoso@utmj.ac.id

ABSTRAK

Media sosial yang sering digunakan oleh pengguna di Indonesia dalam platforme twitter,youtube dan instagram. Dalam penggunaannya twitter, youtube dan instagram digunakan sebagai platform yang membahas tentang opini publik, hiburan dan trending topik didunia salah satu perbincangan pada awal tahun 2020 yakni dihapusnya Ujian Nasional (UN) oleh Kementerian Pendidikan dan Kebudayaan Republik Indonesia (Mendikbud RI). Opini dan Sentimen pengguna di twitter, youtube dan instagram pun sangat beragam, ada yang termasuk kedalam sentimen positif dan ada juga sentimen negatif. Untuk memilah mana yang termasuk kedalam sentimen positif dan negatif diperlukan sebuah rangkaian proses, salah satu proses yang dapat digunakan yakni data mining. Pada penelitian ini akan membahas tentang sentimen pengguna twitter, youtube dan instagram dalam kasus penghapusan UN dengan menggunakan algoritma K-Nearest Neighbor. Pengujian dilakukan menggunakan k-Fold Cross Validation untuk diperoleh nilai akurasi (accuracy), tabel Confusion Matrix dan Area Under Curve. Hasil pengujian diperoleh nilai akurasi 75,45%.

Kata Kunci: *Analisis Sentimen Twitter, Youtube dan Instagram dengan metode K-Nearest Neighbor Penghapusan UN*

ABSTRACT

Social media that is often used by users in Indonesia on twitter, youtube and instagram platforms. In its use twitter, youtube and instagram are used as platforms that discuss public opinion, entertainment and trending topics in the world, one of the discussions in early 2020 was the abolition of the National Examination (UN) by the Ministry of Education and Culture of the Republic of Indonesia (Mendikbud RI). User opinions and sentiments on twitter, youtube and instagram are also very diverse, some are included in positive sentiments and there are also negative sentiments. To sort out which ones are included in positive and negative sentiments, a series of processes is needed, one of the processes that can be used is data mining. This study will discuss the sentiments of twitter, youtube and instagram users in the case of UN deletion using the K-Nearest Neighbor algorithm. The test was carried out using k-Fold Cross Validation to obtain accuracy values, Confusion Matrix tables and Area Under Curve. The test results obtained an accuracy value of 75.45%.

Keywords: *Sentiment Analyst Twitter, Youtube and Instagram with K-Nearest Neighbor UN Removal method*

I. PENDAHULUAN

Perkembangan teknologi saat ini sangat berkembang pesat, salah satunya perkembangan internet. Perkembangan Internet yang begitu cepat telah mengubah banyak aspek dalam proses komunikasi data komputer. Dengan hadirnya internet semakin banyaknya platform-platform yang bisa kita jumpai termasuk platform media sosial.

Media sosial berasal dari kata media dan sosial. Media merupakan alat atau sarana komunikasi seperti koran, radio, televisi, film, poster dan spanduk. Sedangkan sosial merupakan berkenaan dengan masyarakat, memperhatikan kepentingan umum (suka menolong, menderma dan sebagainya). Menurut Mandibergh berpendapat bahwa media sosial adalah media yang mewadahi kerjasama diantara pengguna yang menghasilkan konten (user generated content).

Salah satu platform media sosial yang populer khususnya di Indonesia yakni Twitter, Youtube dan

Instagram. Dalam penghapusan UN tahun 2020 ini pun menuai berbagai opini para pengguna Twitter, Youtube dan Instagram ada yang mendukung kebijakan tersebut dan ada yang tidak mendukung serta tidak sedikit pula ada yang menjadikannya sebagai bahan candaan. Penghapusan UN tersebut dinilai aneh khususnya bagi para calon lulusan karena UN seolah sudah menjadi tradisi turun temurun untuk menentukan lulus atau tidaknya di sekolah tersebut.

Melihat dari hal tersebut, maka dibutuhkan sebuah proses untuk menentukan mana komentar yang merupakan sentimen positif dan mana yang merupakan sentimen negatif. Data mining merupakan salah satu proses yang dapat digunakan untuk mengolah data komentar sentimen para pengguna twitter, Youtube dan Instagram khususnya dalam kasus penghapusan UN untuk mengetahui opini publik yang ada di Twitter, Youtube dan Instagram guna sebagai bahan pertimbangan atau evaluasi untuk para pembuat kebijakan.

Penelitian untuk menentukan analisis sentimen twitter,youtube dan instagram dalam komentar masyarakat menggunakan Algoritma K-Nearest Neighbor. Hasil dari pembahasannya, di dapatkan akurasi sebesar 75,45% dan AUC sebesar 0,733%.

Maka dari penelitian sebelumnya Pada penelitian ini akan membahas tahapan yang dilalui untuk melakukan proses analisis sentimen pada media sosial twitter,youtube dan instagram terhadap penghapusan Ujian Nasional.

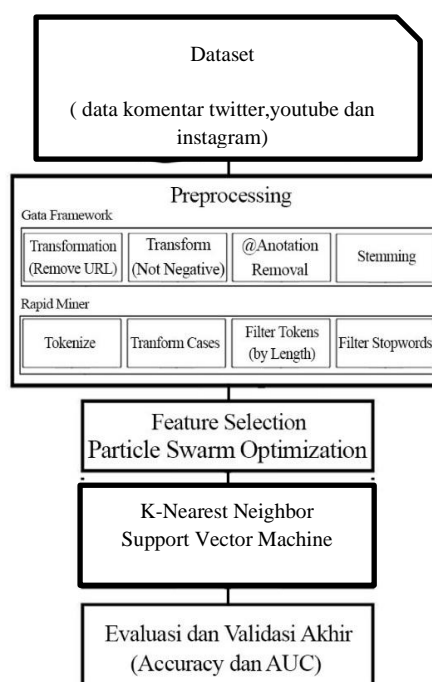
Pada penelitian ini, peneliti menggunakan K-Nearest Neighbor serta *Particle Swarm Optimization* (PSO) sebagai seleksi fitur yang diterapkan untuk meningkatkan performa klasifikasi teks.

Melalui penelitian ini diharapkan bisa mendapatkan hasil yang lebih baik dari penelitian yang sudah dilakukan sebelumnya, serta melalui penelitian ini dapat diketahui perbandingan hasil yang lebih baik dengan metode K-Nearest Neighbor

Pada penelitian analisis sentimen ini, dilakukan untuk melihat sebuah opini seseorang atau masyarakat yang ditujukan kepada penghapusan ujian nasional, sehingga dihasilkan opini yang dapat masuk ke kategori opini negatif ataupun opini positif. Dengan adanya penilaian maupun pandangan masyarakat yang disampaikan pada media sosial, maka peneliti berinisiatif membuat penelitian ini, peneliti melihat belum ada penelitian tentang opinion mining sebelumnya yang mengangkat permasalahan tersebut, sehingga kedepannya diharapkan dapat bermanfaat membantu untuk melakukan riset atas opini masyarakat yang mengandung sentimen positif, negatif ataupun netral lainnya.

II. METODOLOGI

Berikut adalah langkah – langkah metode penelitian yang dijelaskan pada gambar dibawah ini :



Gambar 1. Metode Penelitian

A. Pengumpulan dan Pengolahan Data

Penelitian ini menggunakan data yang didapat dari komentar Twitter akun <https://twitter.com/RamliRizal>. Youtube akun <https://youtu.be/zqiFV8Wn8eY> dan instgram akun <https://www.instagram.com/p/B599fcnFxXR> data yang digunakan dalam penelitian ini sebanyak 350 data yang terdiri dari 215 data *review* positif dan 126 data *review* negatif.

Setelah melakukan pengumpulan data, dilakukan proses pengolahan data atau *preprocessing*, tahapan ini mencakup kegiatan membangun data dan dilanjutkan kegiatan pembersihan data agar dapat dilakukan pengelolaan ketahap selanjutnya. Berikut tahapan *preprocessing* data, antara lain :

1. **Remove URL**
Yaitu proses menghilangkan *Uniform Resource Locator* yang ada pada dataset hasil pengumpulan data. Biasanya digunakan untuk merujuk pada *source* atau sumber dari suatu berita/informasi.
2. **Transform (Not Negative)**
Proses mengubah kata – kata yang bermakna negatif yang akan disatukan dengan tanda garis bawah (_).
3. **Anotation Removal**
Menghapus tanda Mention (@) beserta teks yang ada dibelakangnya. Pada media sosial biasanya berfungsi untuk memanggil user/pengguna lainnya.
4. **Stemming**
Tahapan ini berfungsi mengubah kata perkata menjadi sebuah kata dasar, dengan cara

menghilangkan imbuhan, baik awalan ataupun akhiran.

5. **Tokenize**

Memecah sekumpulan karakter atau kalimat menjadi sebuah potongan karakter atau kata – kata sesuai dengan kebutuhan, biasa juga disebut tokenisasi.

6. **Transforms Cases**

Mengubah huruf kapital yang masih ada di dataset menjadi huruf - huruf kecil. Hal ini bertujuan agar terjadi keseragaman text pada model klasifikasi dan tidak terjadi kesalahan pada proses *tokenize*.

7. **Filter Tokens (By Length)**

Menghilangkan kata – kata dengan panjang karakter tertentu, biasanya kata yang memiliki hanya 2 karakter tidak memiliki arti.

8. **Filter Stopword**

Membuang kata – kata yang diabaikan pada sentimen analisis, biasanya yang berupa kata sambung dan kata keterangan.

Preprocessing pertama dilakukan menggunakan gataframework, yaitu pengolahan data teks berbasis web, dengan menggunakan bahasa pemrograman *php*. Gataframework dapat diakses di alamat link www.gataframework.com/textmining. Banyak yang tersedia di Gataframework, antara lain *annotation removal*, *remove url* dan lainnya. Peneliti menggunakan Gataframework karena memiliki keunggulan dalam melakukan proses *Stemming* bahasa Indonesia. Pengolahan data selanjutnya dilakukan menggunakan tools *Rapidminer*.

B. Seleksi Fitur dan Algoritma

Peneliti mengusulkan menggunakan metode pemilihan seleksi fitur *Particle Swarm Optimization* (PSO) yang digunakan untuk meningkatkan akurasi pada pengklasifikasian algoritma. Algoritma yang akan dijadikan perbandingan atau komparasi pada penelitian ini adalah K-Nearest Neighbor. Peneliti menggunakan K-Nearest Neighbor karena merupakan algoritma mudah dalam data mining melakukan proses pengklasifikasian data.

III. LANDASAN TEORI

A. Analisis Sentimen

Sentiment Analysis, atau yang biasa disebut juga *Opinion Mining*, adalah bidang studi yang bertujuan untuk menganalisis opini, sentimen, penilaian, evaluasi, sikap dan emosi publik terhadap suatu entitas dari produk, pelayanan, suatu permasalahan, organisasi, peristiwa tertentu, topik yang hangat dibicarakan dan atributnya [1].

Ada beberapa macam metode yang bisa digunakan pada analisis sentimen, yaitu antara lain, NB (*Naive Bayes*), *Decision Tree*, KNN (*K-Nearest Neighbor*),

Neural Networks dan juga SVM (*Support Vector Machines*).

B. Particle Swarm Optimization

Particle Swarm Optimization (PSO) pertama kali diusulkan oleh Eberhart dan Kennedy pada tahun 1995, *Particle swarm optimization* (PSO) adalah jenis algoritma kecerdasan yang berasal dari perilaku kawanan burung mencari makan [2]. *Particle Swarm Optimization* dapat diasumsikan sebagai kelompok burung secara mencari makanan disuatu daerah. Burung tersebut tidak tahu dimana makanan tersebut berada, tapi mereka tahu seberapa jauh makanan itu berada, strategi terbaik untuk menemukan makanan tersebut adalah dengan mengikuti burung yang terdekat dari makanan. *Particle swarm optimization* banyak digunakan untuk memecahkan masalah optimasi, serta sebagai masalah seleksi fitur [3].

Menurut Basari bahwa *Particle Swarm Optimization* merupakan teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter. Dalam *Particle Swarm Optimization* terdapat beberapa teknik untuk melakukan pengoptimasian diantaranya meningkatkan bobot atribut (*attribute weight*) terhadap semua atribut atau variabel yang dipakai, menseleksi atribut (*attribute selection*) dan *feature selection* [4].

C. Algoritma K-Nearest Neighbor

Algoritma K-Nearest Neighbor (KNN) merupakan algoritma machine learning sederhana dan mudah diterapkan yang dapat digunakan untuk menyelesaikan masalah klasifikasi dan regresi. Algoritma ini termasuk dalam jenis supervised learning. Metode yang bersifat non-parametric memiliki makna bahwa metode tersebut tidak membuat asumsi apa pun tentang distribusi data yang mendasarinya. Dengan kata lain, tidak ada jumlah parameter atau estimasi parameter yang tetap dalam model, terlepas data tersebut berukuran kecil ataupun besar.

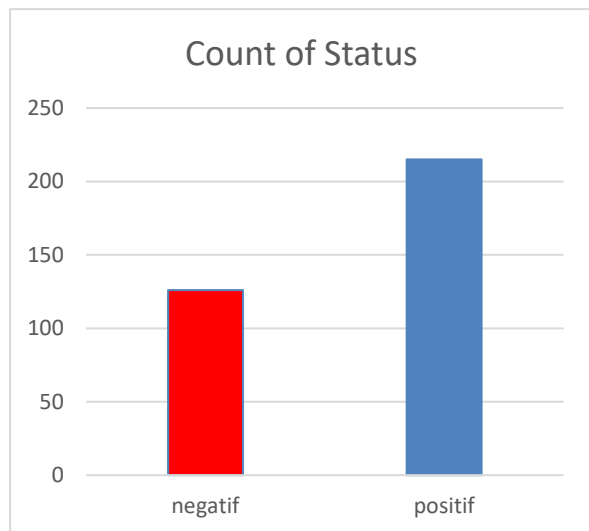
Algoritma non-parametric seperti KNN menggunakan sejumlah parameter yang fleksibel, dan jumlah parameter seringkali bertambah seiring data yang semakin banyak. Algoritma non-parametric secara komputasi lebih lambat, tetapi membuat lebih sedikit asumsi tentang data [5].

IV. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Hasil pengumpulan data yang didapat dari komentar video youtube mengenai peresmian pabrik mobil esemka, peneliti mengumpulkan 288 data opini pengguna youtube. Kemudian data tersebut kumpulkan dalam format file xls, setelah itu data dikelompokkan menjadi 2 bagian yaitu data positif dan data negatif.

Pengelompokan data ini dilakukan oleh beberapa koresponden, sehingga kemudian dihasilkan 180 opini Positif dan 108 opini negatif.



Gambar 2. Statistik hasil pengumpulan data.

B. Pengolahan Data

Setelah pengumpulan data dan membagi data menjadi komentar positif dan komentar negatif, selanjutnya dilakukan proses pengolahan data atau *preprocessing*. Berikut ini adalah tahapan *preprocessing* :

1. Remove URL

Yaitu proses menghilangkan *Uniform Resource Locator* yang ada pada dataset hasil pengumpulan data. Berikut contoh perbedaan sebelum dan sesudah proses.

Tabel 2. Hasil *Remove URL*.

Sebelum	Sesudah
Banyak masyarakat yang bilang di akun https://youtu.be/iYPELEaHi3U untuk melaupkan kekesalannya dan kesenangannya jika un dihapus	Banyak masyarakat yang bilang di akun untuk melaupkan kekesalannya dan kesenangannya jika un dihapus

2. Transform (Not Negative)

Proses mengubah kata – kata yang bermakna negatif yang akan disatukan dengan tanda garis bawah (_).

Tabel 3. Hasil *Transform Negative*.

Sebelum	Sesudah
Murid senang bukan bersedih jika ujian nasional dihapus	Murid senang bukan_bersedih jika ujian nasional dihapus

3. Anotation Removal

Menghapus tanda *Mention* (@) beserta teks yang ada dibelakangnya.

Tabel 4. Hasil *Anotation Removal*.

Sebelum	Sesudah
Ada beberapa murid merasa sedih Cuma beberapa @murid yang memilih untuk diam saja	Ada beberapa murid merasa sedih Cuma beberapa murid yang memilih untuk diam saja

4. Stemming

Tahapan ini berfungsi mengubah kata perkata menjadi sebuah kata dasar, dengan cara menghilangkan imbuhan, baik awalan ataupun akhiran.

Tabel 5. Hasil *Stemming*.

Sebelum	Sesudah
sangat setuju sekali ujian nasional dihapus terlalu banyak oknum yang cari ujian nasional ujian nasional dari pelaksanaan ujian nasional adanya ujian nasional tidak menjamin sumber daya manusianya bagus hanya mengejar nilai bukan ilmu hafal	tuju uji nasional hapus oknum cari uji nasional uji nasional pelaksanaan uji nasional uji nasional jamin sumber daya manusia bagus kejar nilai ilmu hafa

5. Tokenize

Memecah sekumpulan karakter atau kalimat menjadi sebuah potongan karakter atau kata – kata sesuai dengan kebutuhan, biasa juga disebut tokenisasi.

6. Transforms Cases

Mengubah huruf kapital yang masih ada di dataset menjadi huruf - huruf kecil. Hal ini bertujuan agar terjadi keseragaman text pada model klasifikasi dan tidak terjadi kesalahan pada proses *tokenize*.

Tabel 6. Hasil *Transforms Cases*.

Sebelum	Sesudah
anak jaman sekarang diajak diskusi yang ada malah ruang curhat, ujian nasional tolong jangan dihapus.	anak jaman sekarang diajak diskusi yang ada malah ruang curhat ujian nasional tolong jangan dihapus

7. Filter Tokens (By Lenght)

Menghilangkan kata – kata dengan panjang karakter tertentu, biasanya kata yang memiliki hanya 2 karakter tidak memiliki arti.

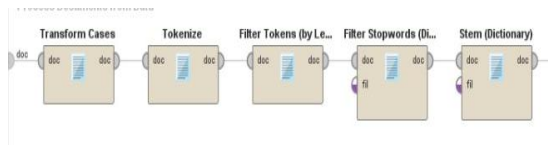
8. Filter Stopword

Membuang kata – kata yang diabaikan pada sentimen analis, biasanya yang berupa kata sambung dan kata keterangan.

Tabel 7. Hasil *Filter Stopword*.

Sebelum	Sesudah
dulu sekolah diajarkan cara membaca menulis demi mengejar yang namanya nilai bagus ketika dewasa hanya berfikir bagaimana cara bertahan hidup di dunia yang semakin keras	sekolah ajar baca tulis kejar nama nilai bagus dewasa berfikir tahan hidup dunia keras

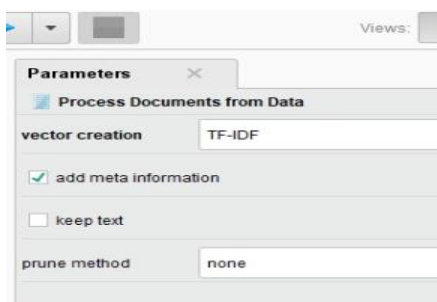
Berikut gambar proses pengolahan data pada *tools rapidminer*.



Gambar 3. Proses Pengolahan Data Rapidminer.

Data hasil *preprocessing* yang berupa kata akan diubah ke dalam bentuk angka dengan dilakukan proses pembobotan kata yang bertujuan untuk menghitung bobot pada masing-masing kata yang akan digunakan sebagai fitur, semakin banyak dokumen yang akan diproses maka semakin banyak fitur. Pada tahapan ini terdapat dua bagian proses yaitu *TF* (*Term Frequency*) dan *IDF* (*Inverse Document Frequency*),

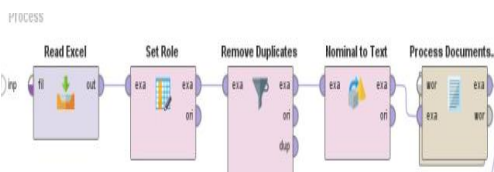
TF adalah jumlah kemunculan tiap kata pada sebuah dokumen semakin banyak kata muncul pada tiap dokumen maka semakin besar nilai *TF*. *Df* adalah jumlah nilai dokumen pada tiap kata yang berbanding terbalik yaitu apabila suatu kata jarang muncul pada sebuah dokumen maka nilai *IDF* lebih besar daripada kata yang sering muncul.



Gambar 4. Proses *Tf Idf*.

C. Fitur seleksi *Particle Swarm Optimization*.

Peneliti mengusulkan menggunakan *Particle Swarm Optimization* untuk meningkatkan akurasi dari pengklasifikasi Algoritma K-Nearest Neighbor. Penelitian ini nantinya menghasilkan akurasi dan nilai *AUC* dengan menggunakan aplikasi *RapidMiner* untuk hasil evaluasi.



Gambar 5. Model Pengujian *PSO* dengan *RapidMiner*.

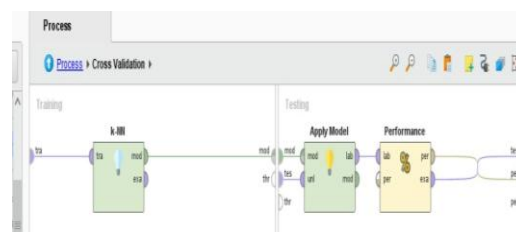
D. Proses Validasi Algoritma.

Selanjutnya adalah penggunaan algoritma *Naive Bayes* dan *Support Vector Machine* untuk melakukan klasifikasi data yang dihubungkan dengan pengujian *10-fold cross validation* dimana proses ini untuk mengevaluasi proses kerja algoritma tersebut dengan membagi data secara acak ke dalam *10 fold* untuk mendapatkan 10 data yang sama, kemudian data tersebut digunakan *9 fold* untuk data *training* dan *1 fold* untuk data *testing*.



Gambar 6. Proses *Cross Validation* dengan *Rapidminer*.

Tahapan selanjutnya adalah proses Evaluasi, ialah mengevaluasi kinerja terhadap permodelan dan perhitungan yang telah digunakan dengan menerapkan algoritma K-Nearest Neighbor yang ditambahkan *Particle Swarm Optimization* agar dihasilkan peningkatan akurasi pada penghitungan. Untuk mengestimasi performa dari model algoritma yang telah dipilih digunakan *Cross Validation* sehingga dapat menghasilkan nilai akurasi (Proses *Cross Validation* dapat dilihat pada Gambar 6). Dibawah ini adalah model pengujian K-Nearest Neighbor. dengan menggunakan *tools Rapidminer*.



Gambar 7. Model Pengujian K-Nearest Neighbor dengan *RapidMiner*

E. Evaluasi

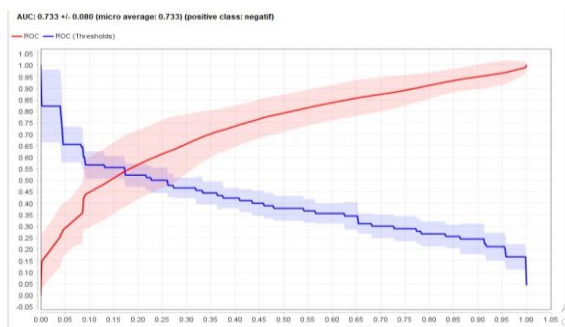
Penelitian ini menggunakan algoritma K-Nearest Neighbor untuk melakukan analisis sentimen. Kemudian dalam menentukan evaluasinya peneliti menggunakan *Accuracy* dan *AUC* (*Area Under Curve*).

	True Positive	True Negative	
Pred Positive	206	65	76.01%
Pred Negative	18	49	73.13%
	91.96%	42.98%	

Pada tabel 8 *Confusion Matrix*

Dapat dilihat, sebanyak 206 data diprediksi *class* positif ternyata sesuai, yaitu masuk kedalam *class* positif, sebanyak 18 data yang diprediksi *class* negatif

ternyata termasuk kedalam prediksi *class* positif. Dan sebanyak 65 data yang diprediksi *class* positif ternyata masuk dalam *class* negatif, kemudian 49 data di prediksi *class* negatif sesuai yaitu termasuk kedalam prediksi *class* negatif.



Gambar 9. Grafik Area Under Curve (AUC) KNN

Berikutnya adalah pengujian yang sudah dilakukan dengan menggunakan *dataset* sebanyak 350 komentar opini pengguna *Youtube, Instagram dan Twitter* mengenai komentar masyarakat tentang penghapusan ujian nasional maka hasil Akurasi Algoritma *Support Vector Machine* tanpa menggunakan Fitur Seleksi *Particle Swarm Optimization* yaitu sebesar 75,45 %, sedangkan untuk nilai *AUC* sebesar 0.733%.

V. KESIMPULAN

Berdasarkan dari penelitian yang sudah dilakukan, dapat disimpulkan bahwa hasil akurasi dan *AUC* dengan algoritma *K-Nearest Neighbor*. bahkan hasil pengujian pengklasifikasian *AUC* nilai keakuratannya dapat dikategorikan sebagai *Excellent Classification*.

Untuk mendukung penelitian ini, peneliti akan mengembangkan aplikasi opini dari komentar youtube mengenai mobil Esemka untuk mengklasifikasikan opini negatif dan positif menggunakan bahasa pemrograman *php*.

DAFTAR PUSTAKA

- Timesindonesia. “UN Dihapus, Pro atau Kontra?” <https://timesindonesia.co.id/kopi-times/245833/un-dihapus-pro-atau-kontra>
- Ilmudapaty. “Algoritma K-Nearest Neighbor untuk Klasifikasi” <https://ilmudapaty.com/algoritma-k-nearest-neighbor-knn-untuk-klasifikasi/>
- [ejournal.uin. “Analisis Sentimen Penghapusan Ujian Nasional pada Twitter Menggunakan Support Vector Machine dan Naïve Bayes berbasis Particle Swarm Optimization” <https://ejournal.uin-suska.ac.id/index.php/coreit/article/download/9723/pdf>
- Romadloni, N. T., Santoso, I., & Budilaksono, S. (2019). Perbandingan Metode Naïve Bayes, KNN dan Decision Tree Terhadap Analisis Sentimen Transportasi KRL Commuter Line. *ikraith-informatika*, 3(2), 1-9.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.