**JSAI : Journal Scientific and Applied Informatics**
Vol. 06, No. 02, Juni 2023, hal. 247~257
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

# Analisis Sentimen Sepak Bola Indonesia pada Twitter menggunakan K-Nearest Neighbors dan Random Forest

**[1]Dedy Agung Prabowo, [2]Sudianto Sudianto**
[1,2]Fakutlas Informatika, Institut Teknologi Telkom Purwokerto, Indonesia
[1]dedy@ittelkom-pwt.ac.id; [2]sudianto@ittelkom-pwt.ac.id;

**Article Info**

***Article history:***

Received, 2023-06-15
Revised, 2023-06-19
Accepted, 2023-06-30

***Kata Kunci:***

K-Nearest Neighbors
NLP
Random Forest
Sentimen
Timnas Indonesia
Twitter

**ABSTRAK**

Twitter merupakan salah satu media sosial yang paling banyak digunakan saat ini. Twitter memungkinkan pengguna untuk memberikan berita dan komentar terbaru tentang peristiwa yang sedang berlangsung di Dunia. Di Indonesia, laga final piala AFF Suzuki Cup 2020 menjadi perbincangan hangat karena untuk keenam kalinya Indonesia menjadi *runner-up* setelah tahun 2000, 2002, 2004, 2010, dan 2016 penampilan Timnas Indonesia. Dengan banyaknya opini dan kritik yang beredar, membedakan opini positif dan negatif membutuhkan waktu yang lama. Oleh karena itu, diperlukan suatu model analisis sentimen yang dapat mengklasifikasikan opini positif dan negatif sebagai bahan evaluasi bagi Timnas Indonesia di masa yang akan datang. Pada penelitian ini, klasifikasi analisis sentimen menggunakan metode algoritme K-Nearest Neighbors dan Random Forest. Data yang digunakan berasal dari balasan kicauan akun Twitter Joko Widodo terkait ucapan selamat kepada Timnas Indonesia usai bertanding melawan Thailand di AFF Suzuki Cup 2020. Berdasarkan hasil pengujian, akurasi algoritme K-Nearest Neighbors 75% lebih baik daripada algoritme Random Forest dengan akurasi 71%

***Keywords:***

***K-Nearest Neighbors***
***NLP***
***Random Forest***
***Sentiment***
***Timnas Indonesia***
***Twitter***

***ABSTRACT***

*Twitter is one of the most widely used social media today. Twitter allows users to provide the latest news and comments about ongoing events in the World. In Indonesia, the final match of the AFF Suzuki Cup 2020 became a hot topic because, for the sixth time, Indonesia was runner-up after 2000, 2002, 2004, 2010, and 2016 appearances for the Indonesian national team. With so many opinions and criticisms circulating, distinguishing between positive and negative opinions takes a long time. Therefore, a sentiment analysis model is needed that can classify positive and negative opinions as evaluation material for the Indonesian National Team in the future. This study uses the K-Nearest Neighbors and Random Forest algorithm methods in sentiment analysis classification. The data comes from a reply to Joko Widodo's Twitter account regarding congratulations to the Indonesian national team after competing against Thailand at the AFF Suzuki Cup 2020. Based on the test results, the accuracy of the K-Nearest Neighbors algorithm is 75% better than the Random Forest algorithm, with an accuracy of 71%..*



***Penulis Korespondensi:***

Dedy Agung Prabowo,
Program Studi Teknik Informatika,
Institut Teknologi Telkom Purwokerto,
Email: dedy@ittelkom-pwt.ac.id

## 1. PENDAHULUAN

Pada era teknologi yang semakin maju telah hadir berbagai macam aplikasi media sosial, salah satunya adalah Twitter. Twitter merupakan salah satu aplikasi media sosial berbasis *website* dan *mobile* yang sering digunakan sebagai alat komunikasi dan dimanfaatkan sebagai media untuk promosi, komunikasi maupun sarana protes [1]. Twitter memiliki jumlah pengguna aktif paling banyak di Indonesia dengan jumlah mencapai 52% dari seluruh media sosial di Indonesia [2]. Hal

ini menjadikan Twitter sangat terkenal dan disukai para pengguna. Pengguna akan memberikan informasi terbaru atau komentar tentang hal yang sedang menjadi *trending* topik atau topik utama di Dunia. Hal yang sedang menjadi topik utama dan banyak dikomentari oleh pengguna di Indonesia adalah topik tentang performa Tim Nasional (Timnas) Indonesia setelah bertanding melawan Timnas Thailand pada laga final piala AFF Suzuki Cup 2020.

Pada laga final piala AFF Suzuki Cup 2020, Thailand keluar sebagai juara umum, sementara Indonesia untuk ke-6 kalinya berada di posisi kedua setelah pada tahun 2000, 2002, 2004, 2010, dan 2016 duduk di posisi yang sama [3]. Setelah Indonesia mengalami kekalahan melawan Thailand, presiden Joko Widodo membuat sebuah kiriman pada akun Twitter pribadinya. Pada kiriman tersebut Joko Widodo memberikan selamat atas perjuangan Timnas pada ajang piala AFF Suzuki Cup 2020. Hal ini membuat beberapa pengguna Twitter memberikan komentar atau opini tentang performa Timnas pada ajang piala AFF Suzuki Cup 2020. Sebelumnya seperti yang kita ketahui, persepakbolaan di Indonesia sedang mengalami banyak polemik dalam beberapa tahun terakhir ini, contohnya seperti pada kasus pergantian ketua umum PSSI, pergantian pelatih Timnas senior, dan pengaturan skor [4]. Polemik yang terjadi tersebut maka akan banyak menimbulkan opini maupun komentar kritikan. Tweet dari pengguna dapat menjadi evaluasi bagi pengelola Timnas Indonesia kedepannya agar performa Timnas sesuai yang diharapkan oleh para pendukung sepak bola.

Penelitian tentang analisis sentimen pada dokumen Twitter telah banyak dilakukan, salah satunya penelitian analisis sentimen pada Twitter menggunakan metode Random Forest dengan TF-IDF mengenai pelanggan hotel di Purwokerto didapatkan hasil akurasi 87,23% dengan proses *stemming* dan tanpa proses *stemming* menghasilkan akurasi sebesar 87,01% [5]. Pada penelitian lain mengenai analisis sentimen Twitter menggunakan metode K-NN mengenai isu terkait kebijakan pemerintah tentang pembelajaran daring didapatkan hasil akurasi tertinggi pada uji *K*=10 sebesar 84,93% dengan *precision* 87%, *recall* 87%, *measure* 87% dan tingkat kesalahan hanya 0,12% [6]. Akan tetapi penelitian analisis tentang performa Timnas pada ajang piala AFF Suzuki Cup 2020 belum pernah dilakukan pada penelitian sebelumnya.

Berdasarkan permasalahan diatas, maka pada penelitian ini dilakukan analisis sentimen Twitter untuk mengklasifikasikan *tweet* opini dan komentar pada performa Timnas Indonesia pada ajang piala AFF Suzuki Cup 2020. Metode yang digunakan dalam penelitian ini adalah metode K-Nearest Neighbors dan Random Forest dengan label positif dan negatif. Metode K-Nearest Neighbors dan Random Forest merupakan salah satu metode untuk klasifikasi. Pengujian dilakukan berdasarkan hasil klasifikasi dan bertujuan untuk mengetahui tingkat akurasi dari klasifikasi. Hasil klasifikasi diharapkan dapat memudahkan pengguna atau pihak terkait dalam melihat *tweet* yang bernilai positif dan negatif.

## 2. PENELITIAN YANG TERKAIT

### *A. Penelitian Sebelumnya*

Pada penelitian sebelumnya yaitu "Perbandingan Metode K-Nearest Neighbors, Decision Tree dan Naive Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS". Penelitian ini menggunakan sebanyak 1000 data dari media sosial Twitter yang dilakukan *filtering* menjadi 903 data. Penelitian ini menghasilkan: (1) nilai akurasi pada metode K-Nearest Neighbors sebesar 95,58% dengan tingkat *precision* pada prediksi *negative* sebanyak 52.17%, prediksi *positive* sebanyak 0.00% sedangkan pada prediksi *neutral* sebanyak 97.27%; (2) tingkat akurasi pada metode Decision Tree sebesar 96,13% dengan tingkat *precision* pada prediksi *negative* sebanyak 55.00%, prediksi *positive* sebanyak 0.00% dan prediksi *neutral* sebanyak 97.28%; (3) tingkat akurasi pada metode Naive Bayes mencapai 89.14% dengan tingkat precision pada prediksi *negative* sebanyak 16.67%, prediksi *positive* sebanyak 1.63% dan prediksi *neutral* sebanyak 98.40% [7].

Penelitian berikutnya berjudul "*Classification Methods on Sentiment Analysis of Tourists on Airlines on Twitter*". *Dataset* yang digunakan diambil dari data Crowdflower dalam format .csv dari Februari 2015 dan mendapatkan *tweet* sebanyak 14.848 data. Penelitian ini melakukan perbandingan terhadap metode Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree dan Gaussian. Hasil terbaik diperoleh dengan menggunakan metode Random Forest dengan tingkat akurasi tertinggi sebesar 75% [8].

Penelitian selanjutnya adalah "Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF". Penelitian tersebut menganalisis ulasan yang terdapat pada salah satu aplikasi e-tourism di Indonesia, yaitu tripadvisor.co.id. *Dataset* yang digunakan adalah sejumlah 1166 komentar dari berbagai jenis hotel di aplikasi tersebut. Metode TF-IDF digunakan dalam seleksi fitur, sedangkan metode Random Forest digunakan dalam proses klasifikasi. Hasil penelitian tersebut menunjukkan metode Random Forest dapat mencapai akurasi model sebesar 87,23% dengan proses *stemming* dan sebesar 87,01% tanpa proses *stemming* [5].

Penelitian selanjutnya "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter Menggunakan Metode K-Nearest Neighbors dan Pembobotan Jumlah *Retweet*". *Dataset* yang digunakan adalah opini masyarakat yang ada pada media sosial Twitter sejumlah 400 data. Nilai *k*=3 merupakan nilai *k* optimal yang digunakan dalam proses klasifikasi menggunakan K-Nearest Neighbors. Hasil klasifikasi diperoleh akurasi sebesar 83,33% dengan melakukan penggabungan pembobotan tekstual dan pembobotan non tekstual [9].

Penelitian selanjutnya "Analisis sentimen pada *tweet* Twitter tentang vaksin COVID-19 menggunakan algoritme klasifikasi KNN". *Dataset* yang digunakan yaitu Sepuluh ribu (10.000) data tweet dari setiap *hastag* vaksin, yaitu *#Pfize*r, *#Moderna*, dan *#AstraZeneca*. Menggunakan *Natural Language Processing* untuk memproses data *tweet* dan K-Nearest Neighbors sebagai model klasifikasi. Hasil dari penelitian mendapatkan sentimen pada vaksin *Pfizer* 47.29% positif, 37.5% negatif dan 15.21% netral. Sentimen pada vaksin *Moderna* 46.16% positif, 40.71% negatif, dan netral 13.13%. Sentimen pada vaksin *AstraZeneca* 40.08% positif, 40.06% negatif, 13.86% netral [10].

Penelitian selanjutnya "Pendekatan Random Forest untuk Analisis Sentimen dalam Bahasa Indonesia". *Dataset* pada penelitian tersebut adalah 386 ulasan berbahasa Indonesia yang diambil dari *FemaleDaily*. Penelitian tersebut menggunakan algoritme Random Forest untuk proses klasifikasi dan mengujinya dengan beberapa metode pembobotan seperti Binary TF, Raw TF, Logarithmic TF, dan TF-IDF. Hasil dari penelitian menunjukkan bahwa analisis sentimen menggunakan Random Forest menghasilkan performa yang baik dengan rata-rata skor OOB (*out-of-bag*) sebesar 0,829. Selain itu penelitian tersebut juga menunjukkan bahwa semua metode pembobotan yang digunakan tidak memiliki efek yang besar untuk analisis sentimen menggunakan Random Forest [11].

Penelitian selanjutnya "Teknik Deep Learning Analisis Sentimen untuk Basis Data Twitter". Dataset tersebut didapatkan dari API database Twitter dengan menggunakan *tools* Hadoop secara *realtime*. Penelitian tersebut menggunakan algoritme Deep Learning dengan tahapan *Preprocessing, Dense Embedding, Hidden Layer* dan klasifikasi sentimen. Hasil penelitian ini menunjukkan bahwa teknik Deep Learning dapat digunakan untuk analisis sentimen pada data Twitter terutama pada data yang jumlahnya sangat banyak [12].

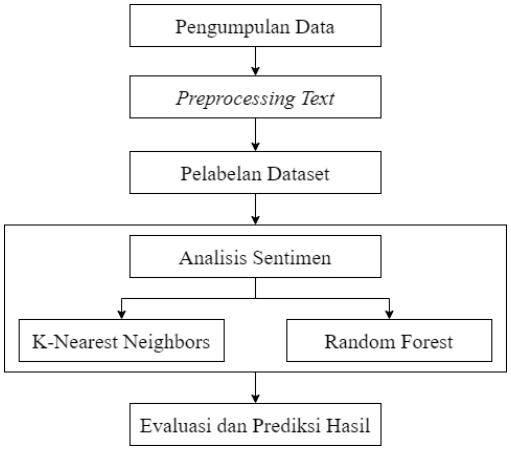
### *B. Dasar Teori*

K-Nearest Neighbors atau yang disingkat dengan KNN merupakan salah satu metode klasifikasi data yang sering dipakai. KNN melakukan klasifikasi data dengan cara menghitung kedekatan jarak antar data. Algoritme KNN mengklasifikasikan data baru dengan menentukan banyaknya nilai dari *k* tetangga terdekat, dimana *k* sebaiknya merupakan angka ganjil positif [13]. Penentuan dekat atau jauhnya jarak dihitung dengan besaran jarak. Besaran jarak yang digunakan adalah *Euclidean Distance*.

Random Forest merupakan metode pengembangan dari metode sebelumnya yaitu *Classification and Regression Tree* (CART). Metode Random Forest menggabungkan banyak pohon (*tree*) untuk membuat klasifikasi, berbeda dengan *decision tree* yang hanya menggunakan *single tree*[27]. Vote terbanyak dari masing-masing *tree* digunakan oleh metode ini untuk menentukan klasifikasi [14].

## 3. METODE PENELITIAN

Dalam melakukan analisis sentimen, diagram alur penelitian dapat dilihat pada Gbr 1.




Gbr. 1 Diagram alur penelitian

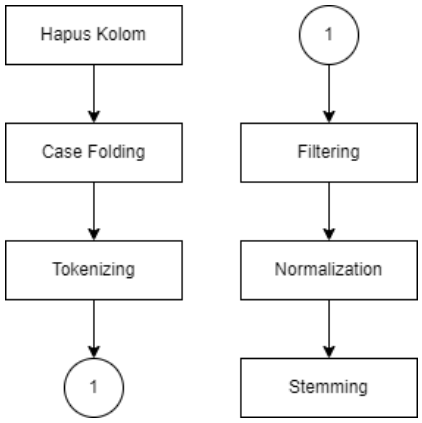
*a) Pengumpulan Data*

Data penelitian berasal dari balasan pada *tweet* akun Twitter Joko Widodo dengan nama akun @jokowi dengan *tweet* seperti yang ditunjukkan pada Gbr 2. Data balasan yang diambil adalah 748 data. Data diambil dengan menggunakan proses *crawling* data dengan menggunakan Bahasa Pemrograman *Python* dalam rentang waktu pada tanggal 1-1-2022 sampai 2-1-2022. Hasil *crawling* data Twitter didapatkan data *username*, tanggal publikasi dan *tweet* yang dikirimkan.


Gbr 2. *Tweet* acuan pada akun Twitter Joko Widodo

*b) Preprocessing Text*

Data yang telah *didapatkan* selanjutnya dilakukan *text preprocessing*, yaitu dengan menghapus kolom yang tidak diperlukan pada tahapan implementasi algoritme, kemudian dilakukan *lower case folding* dengan mengubah setiap huruf yang ada menjadi huruf kecil, selanjutnya dilakukan *tokenizing*, *filtering* (*stopwords removal)*, *normalization* dan terakhir *stemming* [15]–[19].




Gbr 3. Tahapan *preprocessing*

*c) Pelabelan Dataset*

Data yang telah dilakukan proses *preprocessing text* menjadi 684 data yang kemudian dilakukan pelabelan menjadi dua kategori yaitu kategori positif dengan label 1 dan kategori negatif dengan label 0. Pada pemrosesan pelabelan dari 684 data didapatkan data dengan label 1 (positif) sebanyak 464 data dan data dengan label 0 (negatif) sebanyak 220 data.

*d) Analisis Sentimen*

Data yang sudah dilakukan *preprocessing* kemudian dijadikan model dengan dilakukan perhitungan pembobotan tiap data dengan menggunakan perhitungan TF-IDF. Data yang telah dilakukan pembobotan kemudian dilakukan klasifikasi dengan metode K-Nearest Neighbors dan Random Forest untuk kemudian dilakukan analisis hasil. Pada proses sentimen analisis ini dilakukan klasifikasi dan pembuatan model yang sebelumnya dilakukan pembagian data pelatihan dan data pengujian sebanyak 80:20 [20]–[23].

*e) Evaluasi dan Prediksi Hasil*

Evaluasi hasil klasifikasi dilakukan dengan menggunakan *confusion matrix. Confusion matrix* akan memberikan informasi berupa perbandingan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi sebenarnya. *confusion matrix* memiliki bentuk tabel dengan 4 kombinasi berbeda yang direpresentasikan dengan istilah-istilah seperti: *True Positive* (TP), *True Negative* (TN), *False Positive* (TF) dan *False Negative* (FN) [24]–[26].

TABEL I. *CONFUSION MATRIX*

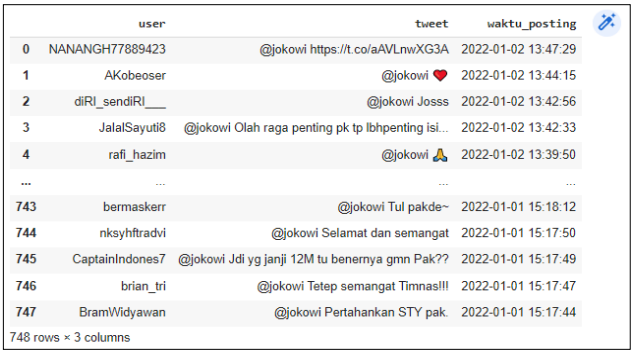
| **Klasifikasi** | | **Prediksi** | |
|---|---|---|---|
| | | *Positive* | *Negative* |
| Aktual | *Positive* | *True Positive (TP)* | *False Negative (FN)* |
| | *Negative* | *False Positive (FP)* | *True Negative (TN)* |

## 4. HASIL DAN PEMBAHASAN

a) Pengumpulan Data

*Dataset* penelitian ini merupakan teks komentar dalam bahasa Indonesia yang mengandung kritik, saran maupun opini. Data penelitian diperoleh dari data komentar yang di-*crawling* dari *reply tweet* pada akun @jokowi terkait Timnas Indonesia setelah bertanding pada babak final piala AFF cup 2020 melawan Thailand. Pengumpulan data dapat dilihat pada Gbr 4.

**JSAI : Journal Scientific and Applied Informatics**
Vol. 06, No. 02, Juni 2023, hal. 247~257
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085



| | user | tweet | waktu_posting |
|---|---|---|---|
| 0 | NANANGH77889423 | @jokowi https://t.co/aAVLnwXG3A | 2022-01-02 13:47:29 |
| 1 | AKobeoser | @jokowi ❤️ | 2022-01-02 13:44:15 |
| 2 | diRI_sendiRI___ | @jokowi Josss | 2022-01-02 13:42:56 |
| 3 | JalalSayuti8 | @jokowi Olah raga penting pk tp lbhpenting isi... | 2022-01-02 13:42:33 |
| 4 | rafi_hazim | @jokowi 🙏 | 2022-01-02 13:39:50 |
| ... | ... | ... | ... |
| 743 | bermaskerr | @jokowi Tul pakde~ | 2022-01-01 15:18:12 |
| 744 | nksyhftradvi | @jokowi Selamat dan semangat | 2022-01-01 15:17:50 |
| 745 | CaptainIndones7 | @jokowi Jdi yg janji 12M tu benernya gmn Pak?? | 2022-01-01 15:17:49 |
| 746 | brian_tri | @jokowi Tetep semangat Timnas!!! | 2022-01-01 15:17:47 |
| 747 | BramWidyawan | @jokowi Pertahankan STY pak. | 2022-01-01 15:17:44 |

748 rows × 3 columns

Gbr 4. Data *tweet* hasil *crawling*

*b) Preprocessing Text*

Pada tahapan *preprocessing* dilakukan untuk membersihkan data dari *noise* maupun dari atribut-atribut yang tidak diperlukan pada saat implementasi algoritme. Tahapan *preprocessing* memiliki peran yang sangat penting pada saat pemodelan analisis sentimen karena keadaan data maupun teks tersebut akan berpengaruh pada hasil akurasi. Ada 5 tahapan *preprocessing* data yaitu *case folding*, *tokenizing*, *filtering* (*Stopword Removal*), *normalization* dan *stemming.*

Pada *preprocessing* data, tidak semua atribut dalam data hasil *crawling* digunakan. Sehingga dilakukan penghapusan kolom-kolom (*dropping columns*) yang tidak diperlukan saat implementasi algoritme. Data yang telah dikumpulkan melalui *crawling* data terdapat 3 kolom, yaitu *user*, *tweet* dan waktu_posting. Pada sentimen analisis yang menjadi objek penelitian adalah *tweet* maka kolom *user* dan waktu_posting perlu dihapus sehingga data yang digunakan seperti di bawah ini:

| | tweet |
|---|---|
| 0 | @jokowi https://t.co/aAVLnwXG3A |
| 1 | @jokowi ❤️ |
| 2 | @jokowi Josss |
| 3 | @jokowi Olah raga penting pk tp lbhpenting isi... |
| 4 | @jokowi 🙏 |
| ... | ... |
| 743 | @jokowi Tul pakde~ |
| 744 | @jokowi Selamat dan semangat |
| 745 | @jokowi Jdi yg janji 12M tu benernya gmn Pak?? |
| 746 | @jokowi Tetep semangat Timnas!!! |
| 747 | @jokowi Pertahankan STY pak. |

748 rows × 1 columns

Gbr 5. Hasil *drop* kolom

Data yang sudah dilakukan penghapusan, tahap selanjutnya yaitu *preprocessing* secara berurutan. Tahapan pertama adalah proses *case folding* untuk merubah penggunaan huruf besar menjadi penggunaan huruf kecil. Sehingga tahapan ini akan menghasilkan kata-kata yang seragam bentuknya yaitu bentuk penggunaan huruf kecil. Proses tahapan *case folding* data disajikan pada Tabel II.

TABEL II. HASIL *CASE FOLDING*

| ***Tweets*** | **Hasil *case folding*** |
|---|---|
| @jokowi Sdh gak kaget dr zsman ke zaman kuda gigit besi timnas pssi blm prh masuk juara sbg perwakilan dr asia utk piala dunia | sdh gak kaget dr zsman ke zaman kuda gigit besi timnas pssi blm prh masuk juara sbg perwakilan dr asia utk piala dunia |

Tahapan kedua adalah *tokenizing* dengan menggunakan *library NLTK* dari data hasil *case folding*. Pada tahapan ini dilakukan proses penghapusan nomor (*number removal*), *whitecase removal*, penghapusan tanda baca dan simbol (*punctuation removal*) dan penggunaan fungsi *word_tokenize()* untuk memecahkan *string* atau kalimat menjadi suatu *tokens*. Hasil dari *tokenizing* dapat dilihat pada Tabel III.

TABEL III. HASIL *TOKENIZING*

| **Hasil *case folding*** | **Hasil *tokenizing*** |
|---|---|
| sdh gak kaget dr zsman ke zaman kuda gigit besi timnas pssi blm prh masuk juara sbg perwakilan dr asia utk piala dunia | ['sdh', 'gak', 'kaget', 'dr', 'zsman', 'ke', 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'blm', 'prh', 'masuk', 'juara', 'sbg', 'perwakilan', 'dr', 'asia', 'utk', 'piala', 'dunia'] |

Tahapan ketiga adalah *filtering* atau *stopword removal* dengan menggunakan *corpus NLTK* dan mengambil data hasil dari *tokenizing*. *Stopwords* pada tahapan *filtering* ini menggunakan Bahasa Indonesia dari *library NLTK* dan menggunakan kamus stopwords-id.txt. Hasil dari data *filtering* direpresentasikan pada Tabel IV.

TABEL IV. HASIL *FILTERING*

| **Hasil *tokenizing*** | **Hasil *filtering*** |
|---|---|
| ['sdh', 'gak', 'kaget', 'dr', 'zsman', 'ke', 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'blm', 'prh', 'masuk', 'juara', 'sbg', 'perwakilan', 'dr', 'asia', 'utk', 'piala', 'dunia'] | ['sdh', 'gak', 'kaget', 'dr', 'zsman', 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'blm', 'prh', 'masuk', 'juara', 'sbg', 'perwakilan', 'dr', 'asia', 'utk', 'piala', 'dunia'] |

Tahapan keempat dilakukan data *normalization* dari data hasil *filtering*. *Normalization* ini dilakukan untuk menyeragamkan *term* yang mempunyai makna atau arti yang sama akan tetapi penulisannya berbeda. Hal ini dilakukan akibat dari adanya kesalahan pada penulisan kata, penyingkatan kata maupun bahasa gaul. Pada tahap ini disiapkan kamus normalisasi.xlsx dan dilakukan penambahan isi dari kamus menyesuaikan dengan dataset untuk melakukan *mapping term* yang akan diseragamkan sehingga menghasilkan kata yang disajikan pada Gbr 6.

| | slang | formal |
|---|---|---|
| 2 | woww | wow |
| 3 | aminn | amin |
| 4 | met | selamat |
| 5 | netaas | menetas |
| 6 | keberpa | keberapa |
| 7 | eeeehhhh | eh |
| 8 | kata2nyaaa | kata-katanya |
| 9 | hallo | halo |
| 10 | kaka | kakak |
| 11 | ka | kak |
| 12 | daah | dah |
| 13 | aaaaahhhh | ah |
| 14 | yaa | ya |
| 15 | smga | semoga |

Gbr 6. Kamus normalisasi

Kemudian kamus normalisasi diimplementasikan pada data hasil *filtering*. Hasil dari data *normalization* dapat dilihat di Tabel 5.

TABEL V. HASIL *NORMALIZATION*

| **Hasil *filtering*** | **Hasil *normalization*** |
|---|---|
| ['sdh', 'gak', 'kaget', 'dr', 'zsman', | ['sudah', 'enggak', 'kaget', 'dari', |

| | |
|---|---|
| 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'blm', 'prh', 'masuk', 'juara', 'sbg', 'perwakilan', 'dr', 'asia', 'utk', 'piala', 'dunia'] | 'zaman', 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'belum', 'parah', 'masuk', 'juara', 'sebagai', 'perwakilan', 'dari', 'asia', 'untuk', 'piala', 'dunia'] |

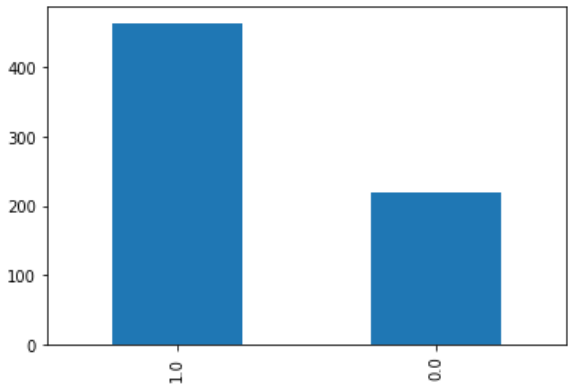
Tahapan terakhir adalah *stemming* menggunakan *library* Sastrawi untuk mengembalikan kata-kata yang sudah dilakukan *normalization* ke bentuk kata dasar. Pada saat melakukan *stemming* menggunakan *library* Swifter untuk mempercepat proses dari *stemming* data. Hasil dari proses *stemming* dapat dilihat pada Tabel VI.

TABEL VI. HASIL *STEMMING*

| **Hasil *normalization*** | **Hasil *stemming*** |
|---|---|
| ['sudah', 'enggak', 'kaget', 'dari', 'zaman', 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'belum', 'parah', 'masuk', 'juara', 'sebagai', 'perwakilan', 'dari', 'asia', 'untuk', 'piala', 'dunia'] | ['sudah', 'enggak', 'kaget', 'dari', 'zaman', 'zaman', 'kuda', 'gigit', 'besi', 'timnas', 'pssi', 'belum', 'parah', 'masuk', 'juara', 'bagai', 'wakil', 'dari', 'asia', 'untuk', 'piala', 'dunia'] |

*c) Pelabelan Dataset*

Data yang telah dilakukan *preprocessing* kemudian masuk ke dalam tahapan pelabelan. Data hasil *preprocessing* tersebut kemudian dilakukan penghapusan pada data yang kosong. Sehingga data yang didapatkan sebanyak 684 data. Pelabelan data dilakukan secara manual dengan hanya melakukan pelabelan pada data hasil dari *preprocessing* dan membedakan balasan yang mengandung kata negatif ditandai dengan label 0 dan positif ditandai dengan label 1. Data yang telah dilakukan pelabelan sebanyak 464 data dengan label 1 dan 220 data dengan label 0. Data yang telah dilabeli inilah yang nantinya akan diolah pada tahapan pembuatan model klasifikasi dengan K-Nearest Neighbors dan Random Forest.




Gbr 7. Jumlah data label 1 dan label 0

*d) Analisis Sentimen*

Tahap pertama yaitu melakukan pembobotan kata menggunakan algoritme TF-IDF yang digunakan untuk data latih (*training*) dan data uji (*testing*). Tahap kedua yaitu pelatihan dan pengujian data. Pada tahap ini membagi data latih (*training*) 80% dan data uji (*testing*) 20% dari jumlah 684 data. Kemudian membuat fungsi "print*"* hasil untuk menampilkan judul, *classification report*, akurasi dan data grafik akurasi dari total data split. Tahap terakhir dari *training* dan *testing* data yaitu membuat fungsi "*klasifikasi data*". Fungsi ini digunakan untuk membuat perulangan hasil *classification report* dengan menggunakan model selection.train_test_split pada *range split* data.

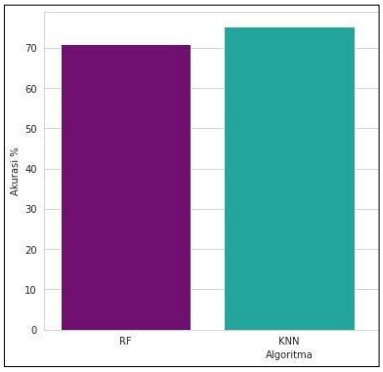
Tahap ketiga yaitu model klasifikasi menggunakan algoritme K-Nearest Neighbors dan Random Forest. Tahapan pertama membuat fungsi "total split" dengan memanggil variabel *dataset* yang akan dilakukan *training* dan *testing* kemudian menentukan *range*

*split*. Pada tahapan ini menggunakan range 0-10 dan memasukkannya ke fungsi klasifikasi data. Fungsi dari algoritme Random Forest dan K-Nearest Neighbors dipanggil untuk membuat model klasifikasi. Hasil yang didapatkan berupa *classification report*, akurasi dan data grafik dari *range* yang telah ditentukan yaitu 0-10. Tahapan terakhir adalah memanggil hasil yang memiliki akurasi tertinggi dan memasukkannya ke dalam fungsi klasifikasi data.

Tahap terakhir adalah membuat grafik *bar plot* menggunakan *library seabron* untuk menampilkan perbandingan akurasi dari algoritme K-Nearest Neighbors dan Random Forest. Visualisasi hasil perbandingan akurasi algoritme K-Nearest Neighbors dan Random Forest yang ditampilkan pada Tabel VII dan Gbr 8.

TABEL VII. HASIL PERBANDINGAN METODE K-NEAREST NEIGHBORS DAN RANDOM FOREST

| Algoritme | Akurasi (%) |
|---|---|
| KNN | 75,18 |
| RF | 70,80 |




Gbr 8. Hasil perbandingan metode K-Nearest Neighbors dan Random Forest

*e) Evaluasi dan Prediksi Hasil*

Pada tahap evaluasi, dilakukan pengujian model klasifikasi menggunakan *confusion matrix.* Hasil dari *confusion matrix* pada model klasifikasi Random Forest yaitu *accuracy* 71%, *recall* 56%, *precision* 70% *dan f1 score* 54%. Selain itu, hasil dari *confusion matrix* pada model klasifikasi K-Nearest Neighbors yaitu *accuracy* 75%, *recall* 66%, *precision* 73% *dan f1 score* 67%.

TABEL VIII. PERBANDINGAN HASIL *ACCURACY, RECALL, PRECISION DAN F1 SCORE*

| Hasil | KNN (%) | RF (%) |
|---|---|---|
| Akurasi | 75 | 71 |
| *Precision* | 73 | 70 |
| *Recall* | 66 | 56 |
| *F1 Score* | 67 | 54 |

Hasil evaluasi pada analisis sentimen balasan *tweet* tentang Timnas pada akun Twitter Joko Widodo menunjukan bahwa K-Nearest Neighbors lebih unggul dari pada Random Forest (Tabel VIII). Berdasarkan penggunaannya baik K-Nearest Neighbors maupun Random Forest umum digunakan untuk model klasifikasi. Akan tetapi pada Random Forest biasanya digunakan untuk *dataset* yang banyak, sedangkan K-Nearest Neighbors adalah klasifikasi yang paling sederhana yaitu mengelompokkan suatu objek dengan dasar perhitungan jarak terdekat dengan objek lainnya. Hasil akurasi juga bisa dipengaruhi karena kurangnya data dan ukuran pelabelan data yang tidak seimbang.

## 5. KESIMPULAN

Berdasarkan hasil analisis sentimen yang telah dilakukan. Algoritme K-Nearest Neighbors dan Random Forest dapat digunakan untuk menganalisis sentimen pengguna Twitter dengan memanfaatkan data dari *reply tweet* pada akun Jokowi terkait Timnas Indonesia setelah bertanding pada babak final piala AFF Cup 2020 melawan Thailand. Analisis sentimen terhadap data Twitter terkait Timnas Indonesia setelah bertanding pada babak final piala AFF Cup 2020 melawan Thailand menghasilkan nilai akurasi sebesar 75%, *precision* sebesar 73%, *recall* sebesar 66%, dan *f1-score* sebesar 67% menggunakan algoritme K-Nearest Neighbors. Sedangkan algoritme Random Forest menghasilkan akurasi sebesar 71%, dengan *precision* sebesar 70%, *recall* sebesar 56%, dan *f1-score* sebesar 54%. Dalam penelitian ini diketahui bahwa algoritme K-Nearest Neighbors lebih unggul dibandingkan algoritme Random Forest

## UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada Jurusan Teknik Informatika, Institut Teknologi Telkom Purwokerto atas dukungan dalam menyelesaikan penelitian ini, serta pihak yang tidak dapat disebut satu per satu

## REFERENSI

[1] A. Wahyuningtyas, I. S. Sitanggang, and H. Khotimah, "Deteksi Spam pada Twitter Menggunakan Algoritme Naïve Bayes," *Jurnal Ilmu Komputer dan Agri-Informatika*, vol. 7, no. 1, pp. 31–40, 2020, doi: 10.29244/jika.7.1.31-40.

[2] F. Fitriana, E. Utami, and H. Al Fatta, "Analisis Sentimen Opini Terhadap Vaksin Covid - 19 pada Media Sosial Twitter Menggunakan Support Vector Machine dan Naive Bayes," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 5, no. 1, pp. 19–25, 2021, doi: 10.31603/komtika.v5i1.5185.

[3] A. Fajriansyah, "Indonesia Jadi 'Runner Up' Keenam Kali," *kompas*, 2022. https://www.kompas.id/baca/olahraga/2022/01/01/piala-aff-2020-Indonesia-raih-runner-up-keenam-kali (accessed Jan. 05, 2022).

[4] J. A. Septian, T. M. Fahrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *JOURNAL OF INTELLIGENT SYSTEMS AND COMPUTATION*, pp. 43–49, 2019.

[5] B. B. Baskoro and S. K. Susanto, Irwan, "Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR)," *Journal of Informatics, Information System, Software Engineering and Applications*, vol. 8106, pp. 21–29, 2021, doi: 10.20895/INISTA.V3.

[6] T. N. P. Dicki Pajri, Yuyun Umaidah, "Implementation of K-Nearest Neighbor ( K-NN ) Algorithm For Public Sentiment Analysis of Online Learning," *Teknik Informatika dan Sistem Informasi*, vol. 15, no. 2, pp. 121–130, 2021, doi: 10.22146/ijccs.65176.

[7] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[8] E. Fitriana Saraswita and D. Palupi Rini, "Classification Methods on Sentiment Analysis of Tourists on Airlines n Twitter," *Sriwijaya Journal of Informatic and Applications*, vol. 2, no. 1, pp. 1–7, 2021.

[9] W. E. Nurjanah, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, vol. 1, no. 12, pp. 1750–1757, 2017.

[10] F. M. J. M. Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 463–470, 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.

[11] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp. 46–50, 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[12] A. M. Gattan, "Deep Learning Technique of Sentiment Analysis for Twitter Database," vol. Vol. 16, N, pp. 184–193, 2022.

[13] E. Laksono, A. Basuki, and F. Bachtiar, "Optimization of K Value in KNN Algorithm for Spam and Ham Email Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 2, pp. 377–383, 2020, doi: 10.29207/resti.v4i2.1845.

[14] M. A. Ghani and A. Subekti, "Email Spam Filtering Dengan Algoritma Random Forest," *IJCIT (Indonesian Journal on Computer and Information Technology*, vol. 3, no. 2, pp. 216–221, 2018.
[15] S. Sudianto, P. Wahyuningtias, H. W. Utami, U. A. Raihan, and H. N. Hanifah, "Comparison Of Random Forest And Support Vector Machine Methods On Twitter Sentiment Analysis ( Case Study : Internet Selebgram Rachel Vennya Escape From Quarantine ) Perbandingan Metode Random Forest Dan Support Vector Machine Pada Analisis Sentimen Twitt," *Jutif*, vol. 3, no. 1, pp. 141–145, 2022.
[16] S. Sudianto, A. D. Sripamuji, I. R. Ramadhanti, R. R. Amalia, J. Saputra, and B. Prihatnowo, "Penerapan Algoritma Support Vector Machine dan Multi-Layer Perceptron pada Klasifikasi Topik Berita," *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, vol. 11, no. 2, pp. 84–91, 2022.
[17] W. Afandi, S. N. Saputro, A. M. Kusumaningrum, H. Ardiansyah, M. H. Kafabi, and S. Sudianto, "Klasifikasi Judul Berita Clickbait menggunakan RNN-LSTM," *Jurnal Pengembangan IT*, vol. 7, no. 2, pp. 85–89, 2022.
[18] S. Sudianto, J. A. Marseli, N. Nugroho, R. W. A. Rumpoko, and Z. Akhmad, "Comparison of Support Vector Machines and K-Nearest Neighbor Algorithm Analysis of Spam Comments on YouTube Covid Omicron," *JTI*, vol. 15, no. 2, pp. 110–118, doi: https://doi.org/10.15408/jti.v15i2.24996.
[19] S. Chandra Ayunda Apta, N. Trivetisia, N. A. Winanti, D. P. Martiyaningsih, T. W. Utami, and S. Sudianto, "Analisis Komparasi Algoritma Machine Learning untuk Sentiment Analysis (Studi Kasus: Komentar YouTube 'Kekerasan Seksual')," *Jurnal Pengembangan IT*, vol. 7, no. 2, pp. 80–84, 2022.
[20] Sudianto, Y. Herdiyeni, A. Haristu, and M. Hardhienata, "Chilli quality classification using deep learning," in *2020 International Conference on Computer Science and Its Application in Agriculture, ICOSICA 2020*, 2020. doi: 10.1109/ICOSICA49951.2020.9243176.
[21] Sudianto, Y. Herdiyeni, and L. B. Prasetyo, "Machine learning for sugarcane mapping based on segmentation in cloud platform," presented at the The 3rd International Conference on Engineering, Technology and Innovative Researches, Purwokerto, Indonesia, 2023, p. 020001. doi: 10.1063/5.0132180.
[22] S. Sudianto, "Analisis Kinerja Algoritma Machine Learning Untuk Klasifikasi Emosi," vol. 4, no. 2, pp. 1027–1034, 2022, doi: 10.47065/bits.v4i2.2261.
[23] R. M. S. Adi and S. Sudianto, "Prediksi Harga Komoditas Pangan Menggunakan Algoritma Long Short-Term Memory ( LSTM )," vol. 4, no. 2, pp. 1137–1145, 2022, doi: 10.47065/bits.v4i2.2229.
[24] T. K. Putri, M. L. Arnumukti, K. Khatimah, E. Zalsabila, and S. Sudianto, "Diabetes Diagnostic Expert System using Website-Based Forward Chaining Method," *Data Science*, vol. 3, no. 1, 2023.
[25] T. Widodo, S. Maghfiroh, S. H. B. Ginting, A. Aryaputra, and S. Sudianto, "Prediction of Covid-19 Cases in Central Java using the Autoregressive (AR) Method," *Data Science*, no. 1, 2023.
[26] A. M. Yusuf, J. I. Chelidivano, T. A. Rizky, Y. Sabikhi, and S. Sudianto, "An Expert System for Diagnosing the Impact of Traffic Accidents using the Forward Chaining Method," *Data Science*, vol. 3, no. 1, 2023.
[27] Kaban, T. M., Astiti, S., & Prabowo, D. A. (2023). Perancangan Aplikasi Pelaporan Harian dengan Design Thinking dan User Experience Questionnaire (UEQ). *JURIKOM (Jurnal Riset Komputer)*, *10*(2), 603-614.