



Penggunaan Algoritma K-Nearest Neighbors (K-NN) dalam Klasifikasi Artikel Clickbait Berbahasa Indonesia

Utilizing the K-Nearest Neighbors (K-NN) Algorithm for Categorizing Clickbait Articles in Indonesian Language

Laila Isyriyah¹, Adi Bayu Permadi², Rakhmad Maulidi^{3*}

^{1,2,3}Informatika, STIKI Malang

¹laila@stiki.ac.id, ²adibayu92@gmail.com, ³maulidi@stiki.ac.id

Abstract

Clickbait is a strategy commonly used to attract readers' attention with promising sensational or intriguing headlines. However, often these clickbait headlines do not correspond to the actual content of the news, resulting in disappointment for the readers. Therefore, this study aims to classify clickbait news headlines in the Indonesian language using the K-Nearest Neighbors (K-NN) method. The purpose of this research is to evaluate the ability of the K-NN method to classify clickbait news headlines in the Indonesian language. Thus, it is expected to provide a better understanding of the effectiveness of this method in identifying clickbait headlines. This study utilizes the K-NN method to classify clickbait news headlines. The data consists of 800 training data and 200 test data. The training and testing processes are conducted by varying the number of neighbors (k) and using various supporting features. The results show that the best performance of the K-NN method is achieved with a number of neighbors $k=11$, yielding an accuracy of 80.5%, Precision of 85%, Recall of 81%, and F-measure of 80%. Testing with 20 new data also resulted in an accuracy rate of 90%. Additionally, several unique words that frequently appear in clickbait headlines are identified, such as "apa" (what), "kenapa" (why), "nih" (here), "alasan" (reason), and "wow". This research contributes to identifying clickbait news headlines in the Indonesian language using the K-NN method. The findings of this study can serve as a reference for further research and provide better insights into how the K-NN method can be applied in classifying clickbait headlines.

Keywords: clickbait detection, article classification, online content analysis, natural language processing, machine learning algorithm, K-Nearest Neighbors

Abstrak

Clickbait merupakan strategi yang sering digunakan untuk menarik perhatian pembaca dengan judul yang menjanjikan sensasi atau rasa penasaran. Namun, sering kali judul-judul clickbait tersebut tidak sesuai dengan isi berita yang sebenarnya, menyebabkan kekecewaan bagi pembaca. Tujuan dari penelitian mengklasifikasikan judul-judul berita clickbait dalam bahasa Indonesia menggunakan metode K-Nearest Neighbors (K-NN). Selain itu tujuan lain dari penelitian ini adalah untuk mengevaluasi kemampuan metode K-NN dalam mengklasifikasikan judul berita clickbait dalam bahasa Indonesia. Dengan demikian, diharapkan dapat memberikan pemahaman yang lebih baik tentang efektivitas metode ini dalam mengidentifikasi judul-judul clickbait. Penelitian ini menggunakan metode K-NN untuk mengklasifikasikan judul-judul berita clickbait. Data yang digunakan terdiri dari 800 data latih dan 200 data uji. Proses pelatihan dan pengujian dilakukan dengan memvariasikan jumlah tetangga (k) dan menggunakan berbagai fitur pendukung. Hasil penelitian menunjukkan bahwa kinerja terbaik metode K-NN tercapai pada jumlah tetangga $k=11$, dengan akurasi 80.5%, Presisi 85%, Recall 81%, dan F-Measure 80%. Pengujian dengan 20 data baru juga menghasilkan tingkat akurasi sebesar 90%. Selain itu, beberapa kata unik yang sering muncul dalam judul clickbait juga diidentifikasi, seperti "apa", "kenapa", "nih", "alasan", dan "wow". Penelitian ini memberikan kontribusi dalam mengidentifikasi judul berita clickbait dalam bahasa Indonesia menggunakan metode K-NN. Hasil penelitian ini dapat menjadi referensi bagi penelitian selanjutnya dan memberikan wawasan yang lebih baik tentang bagaimana metode K-NN dapat diterapkan dalam klasifikasi judul-judul clickbait.

Kata kunci: deteksi clickbait, klasifikasi artikel, analisa content daring, pembelajaran mesin, K-Nearest Neighbors

1. Pendahuluan

Saat ini, memperoleh informasi tidak lagi dianggap sebagai suatu tugas yang menantang berkat ketersediaan internet. Internet telah memfasilitasi proses pencarian informasi secara luas dan fleksibel, menyediakan akses yang mudah di berbagai lokasi dan waktu. Dengan perkembangan yang terus berlanjut, internet telah menjadi platform utama bagi banyak individu dalam upaya mereka mencari beragam jenis informasi. Keberadaan internet memungkinkan akses terhadap informasi dengan bervariasi [1], tidak selalu tersedia dalam format cetak seperti buku, majalah, koran, atau melalui saluran televisi konvensional [2]. Baik melalui situs berita resmi maupun blog pribadi, internet terus menjadi wadah yang berkembang bagi penyediaan informasi di ranah maya.

Pentingnya kebenaran dan ketepatan suatu informasi tidak selalu menjadi prioritas utama bagi sebagian penyedia informasi [3]. Sebaliknya, penyedia informasi sering kali lebih fokus utama pada upaya untuk meningkatkan jumlah pengunjung atau lalu lintas (traffic) pada situs [4], bahkan jika itu berarti mengorbankan kualitas konten yang disampaikan kepada pembaca. Tingginya jumlah pengunjung dianggap dapat meningkatkan pendapatan situs melalui iklan yang ditampilkan [2]. Salah satu strategi yang digunakan oleh pengelola situs untuk menarik perhatian pengunjung dan meningkatkan lalu lintas adalah dengan menerapkan teknik Clickbait.

Clickbait merupakan strategi yang memanfaatkan judul yang berlebihan atau hiperbolik guna menarik perhatian pembaca [2], mendorong mereka untuk mengklik dan membaca artikel tersebut. Judul yang menarik dan terkesan sensasional, pembaca merasa penasaran dan tertarik untuk mendalami konten yang disajikan. Praktik ini dikenal dapat meningkatkan peluang peningkatan jumlah pengunjung situs dibandingkan dengan situs yang tidak menggunakan clickbait [5].

Penggunaan clickbait dapat menjadi alat efektif bagi situs berita untuk menarik lebih banyak pembaca, asalkan tetap memperhatikan dan menyajikan konten berita yang berkualitas dan sesuai [3]. Namun, permasalahannya muncul ketika beberapa situs menyalahgunakan clickbait dengan menyajikan konten berita yang tidak sesuai dengan judulnya, tidak berkualitas, atau bahkan bersifat palsu [6]. Tujuan utama dari tindakan tersebut adalah meningkatkan traffic dan jumlah pengunjung situs [4], yang pada akhirnya berdampak pada peningkatan pendapatan dari iklan yang ditampilkan di situs tersebut.

Penelitian tentang pengembangan/penggunaan algoritma untuk mendeteksi atau mengklasifikasi clickbait sudah banyak dilakukan, umumnya menggunakan algoritma machine learning klasik, seperti Logistic Regression, Naïve Bayes beserta

variannya, Random Forest, Support Vector Machine, Decision Tree [1].

Penelitian dengan pendekatan Deep learning mulai banyak bermunculan seperti Deep Attentive Model [7], Transformers [2], BRET [4]. K-Nearest Neighbors (K-NN) termasuk dalam algoritma machine learning klasik yang banyak digunakan untuk klasifikasi atau prediksi.

Cara kerja algoritma K-NN adalah dengan mengukur jarak antara titik data baru (instance) yang akan diprediksi dengan titik-titik data pelatihan yang ada. Salah satu keunggulan dari K-NN adalah sederhana dalam konseptualisasi dan implementasi, serta mudah diterapkan pada berbagai jenis masalah klasifikasi dan regresi. Namun, algoritma ini memiliki beberapa kelemahan, seperti sensitif terhadap skala fitur, rentan terhadap adanya data pencilan, dan memerlukan penyimpanan yang besar untuk menyimpan data pelatihan secara keseluruhan.

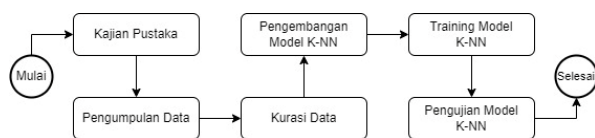
Untuk mengatasi masalah penyalahgunaan clickbait, disarankan untuk melakukan pengelompokan artikel menjadi dua kategori, yaitu clickbait dan non-clickbait. Salah satu metode yang dapat digunakan untuk mempermudah proses klasifikasi adalah metode K-NN. Metode ini digunakan untuk mengklasifikasikan data atau objek ke dalam suatu kelompok berdasarkan data pembelajaran atau sampel latih yang memiliki jarak paling dekat dengan data yang akan diklasifikasikan [1].

Tujuan penelitian ini adalah untuk mengelompokkan judul-judul berita clickbait dengan menggunakan pendekatan metode K-NN. Selain itu, tujuan lainnya adalah untuk menilai efektivitas metode K-NN dalam membedakan judul-judul berita clickbait atau bukan. Pengelompokan judul berita menggunakan dalam penelitian hanya terbatas pada judul berita berbahasa Indonesia.

Kontribusi dalam penelitian ini adalah mengembangkan metode klasifikasi judul berita clickbait menggunakan algoritma K-NN, yang telah terbukti efektif dalam berbagai jenis masalah klasifikasi [1][2]. Penelitian ini berfokus pada penerapan K-NN untuk mengelompokkan judul berita berbahasa Indonesia ke dalam kategori clickbait dan non-clickbait. Dengan demikian, penelitian ini tidak hanya menawarkan solusi praktis untuk mendeteksi clickbait, tetapi juga mengevaluasi keefektifan K-NN dalam konteks bahasa Indonesia, yang masih jarang dibahas dalam literatur. Hasil dari penelitian ini diharapkan dapat membantu meningkatkan kualitas konten di internet dengan mengurangi penyebaran judul-judul yang menyesatkan.

2. Metode Penelitian

Penelitian ini terdiri dari 6 langkah penelitian seperti terlihat pada Gambar 1.



Gambar 1. Metode Penelitian

Tahap pertama yang dilakukan dengan menentukan topik penelitian yakni tentang pendeteksi judul artikel clickbait berbahasa Indonesia. Dilanjutkan pencarian literatur dari jurnal ilmiah di Google Scholar yang terkait topik penelitian dengan menggunakan keyword *clickbait detection*, *article classification*, *online content analysis*, *natural language processing*, *machine learning algorithm*. Hasil pencarian kemudian diseleksi berdasarkan relevansi dengan topik penelitian, dilanjutkan pembuatan literatur review untuk penentuan arah/tujuan penelitian yang akan dilakukan.

Objek pada penelitian ini adalah kumpulan judul artikel atau berita yang diperoleh dari situs penyedia berita online. Data dikumpulkan dari sejumlah situs berita populer berbahasa Indonesia dengan menggunakan teknik web scraping, yakni sebuah Teknik untuk mengambil data dari sebuah website untuk diambil data-data yang hanya diperlukan, yakni hanya judul beritanya saja. Teknik ini dilakukan dengan cara mengekstrak teks [8] dari elemen html dimana judul berita ditulis dalam kode html. Misalnya pada situs detik.com, judul berita ditulis menggunakan tag header1 `

` dan kelas `detail_title`, dengan menggunakan Bahasa pemrograman python dan library BeautifulSoup dengan potongan baris kode berikut, judul berita bisa didapatkan dengan mudah.

```
# Parsing HTML menggunakan BeautifulSoup
soup = BeautifulSoup(html_data, 'html.parser')

# Mengekstrak teks dari elemen h1 dengan kelas 'detail_title'
judul_artikel = soup.find('h1',
class_='detail_title').text.strip()
```

Data-data judul artikel berita yang akan digunakan sebagai objek penelitian ini akan dikumpulkan dari situs berikut bintang.com, brilio.net, detik.com, jawapos.com, kapanlagi.com, kompas.com, merdeka.com, okezone.com, otomania.com, republika.co.id, tempo.co, dan tribunnews.com. Diperlukan penyesuaian kode pengambilan judul berita untuk setiap situs, karena penulisan kode html, tag dan kelas yang berbeda untuk masing-masing situs. Dataset tambahan didapatkan dari Click-ID [9] untuk memberikan variasi data yang lebih bervariasi.

Proses kurasi data untuk penelitian klasifikasi clickbait judul berita berbahasa Indonesia dari beberapa portal web nasional di Indonesia memerlukan beberapa langkah penting. Pertama, identifikasi sumber data dari beberapa portal berita terkemuka di Indonesia, seperti detik.com, kompas.com, atau liputan6.com. Selanjutnya, lakukan seleksi data dengan memilih judul-judul berita yang relevan dengan topik penelitian,

seperti yang berkaitan dengan politik, hiburan, atau olahraga. Setelah itu, lakukan penyaringan data dengan memastikan bahwa judul-judul berita yang dipilih sesuai dengan kriteria clickbait, seperti penggunaan judul yang provokatif atau sensasional untuk menarik perhatian pembaca [10]. Validasi data dilakukan dengan memverifikasi keakuratan dan konsistensi judul-judul berita yang dikumpulkan dari berbagai sumber. Selain itu, normalisasi data juga perlu dilakukan untuk memastikan konsistensi format dan struktur judul berita. Terakhir, dokumentasikan proses kurasi data dengan baik, termasuk sumber data, kriteria seleksi, dan metode penyaringan yang digunakan, untuk memastikan transparansi dan keandalan dalam penelitian klasifikasi clickbait ini [9].

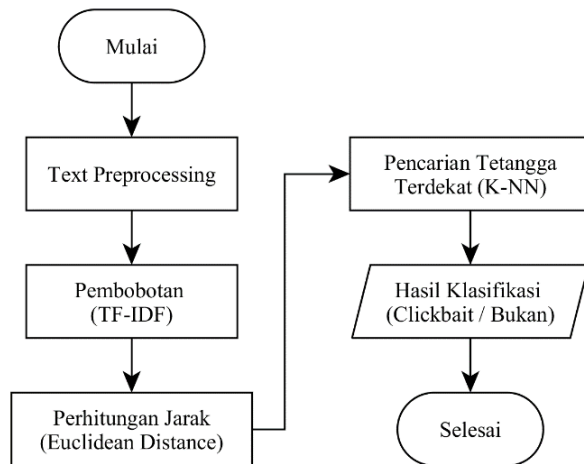
Tabel 1 Kurasi Dataset Judul Berita

No.	Judul Berita	Clickbait
1.	Begini Kondisi Jalan di Sekitar Masjid Istiqlal dalam Aksi 313	Ya
2.	Ini yang Bikin Cassandra Lee Susah Berpaling dari Randy Martin	Ya
3.	Ahok: Kalau Lift Sudah Bunyi, Jangan Memaksa Masuk	Tidak
4.	Sandiaga Uno: PAN Akan Jadi Tulang Punggung Program OK OCE	Tidak
5.	Kevin Sanjaya Sukamuljo/Marcus Fernaldi Gideon Juara Ganda All England 2017	Tidak
6.	Diproteks, Ini Jawaban Nike Soal Produk Hijab	Ya
7.	Ridwan Kamil Diusung Nasdem Sebagai Cagub Jabar	Tidak
8.	Apakah Anda Kurang Minum? Ini Cara Mudah Mengetahuinya	Ya
9.	Angka Pengangguran di Jabar Masih Tinggi Disebut Tak Perhatian Pada Anak, Ini Jawaban Pedas Ussy Sulistiaawaty	Ya

Tabel 1 merupakan contoh kurasi dataset judul berita dengan kolom "Judul Berita" dan "Clickbait". Kolom "Judul Berita" mencantumkan berbagai judul berita, sementara kolom "Clickbait" menunjukkan apakah judul tersebut dianggap sebagai clickbait atau bukan. Dataset ini terdiri dari judul-judul yang diberi label sebagai clickbait (Ya) dan bukan clickbait (Tidak), dan akan digunakan untuk melatih serta menguji model klasifikasi yang dikembangkan.

Proses pengembangan model K-NN dalam penelitian ini melibatkan serangkaian langkah sistematis. Pertama, data yang telah dikumpulkan dari berbagai portal berita berbahasa Indonesia dikurasi dengan hati-hati, memilih judul-judul berita yang mewakili variasi clickbait dan non-clickbait sejumlah 1000 data. Selanjutnya, data ini dibagi menjadi dua subset: satu untuk pelatihan model yakni sejumlah 800 data dan 200 data untuk pengujian. Kemudian, dilakukan pra-pemrosesan data, termasuk tokenisasi teks, pembersihan data, dan pembentukan representasi vektor teks menggunakan teknik TF-IDF (Term Frequency-Inverse Document Frequency) [11]. Selanjutnya, model K-NN dikembangkan dengan menyesuaikan parameter k, yaitu jumlah tetangga terdekat yang akan digunakan untuk klasifikasi [12].

Alur diagram model K-NN pada penelitian ini terdiri dari beberapa tahapan seperti terlihat pada Gambar 2. Tahap pertama adalah text processing; dilanjutkan dengan tahap pembobotan TF-Idf yang merupakan fitur dari model K-NN; selanjutnya adalah tahap perhitungan jarak (*Euclidean Distance*); dilanjutkan dengan proses K-NN dan tahap yang terakhir dari alur diagram model K-NN klasifikasi artikel clickbait.



Gambar 2. Model K-NN

Proses pelatihan model dilakukan dengan menggunakan subset data pelatihan dan pengoptimalan parameter dilakukan melalui validasi silang untuk meningkatkan kinerja model. Akhirnya, model yang telah dilatih dievaluasi menggunakan subset data pengujian untuk mengukur kinerjanya dalam mengklasifikasikan judul-judul berita sebagai clickbait atau non-clickbait [13]. Hasil evaluasi ini digunakan untuk menilai keefektifan dan keakuratan model K-NN dalam konteks klasifikasi clickbait berita dalam bahasa Indonesia.

Proses pengujian model K-NN untuk penelitian klasifikasi clickbait judul berita berbahasa Indonesia merupakan tahap kritis dalam validasi dan evaluasi kinerja model. Setelah model K-NN dikembangkan dan dilatih menggunakan data pelatihan, langkah selanjutnya adalah menguji model tersebut menggunakan subset data pengujian yang sebelumnya tidak digunakan dalam proses pelatihan. Judul-judul berita dalam data pengujian akan diberikan kepada model untuk diklasifikasikan sebagai clickbait atau non-clickbait. Selama proses pengujian, model akan memprediksi label untuk setiap judul berita berdasarkan pola dan fitur-fitur yang telah dipelajari selama proses pelatihan. Hasil prediksi kemudian dibandingkan dengan label sebenarnya dari data pengujian untuk mengukur kinerja model, termasuk akurasi, presisi, recall, dan F1-score [13]. Evaluasi ini membantu peneliti untuk memahami seberapa baik model K-NN dapat membedakan judul-judul clickbait dari non-clickbait dalam konteks berita berbahasa Indonesia.

Selain itu, hasil pengujian juga dapat digunakan untuk mengidentifikasi area di mana model perlu ditingkatkan atau disesuaikan untuk meningkatkan kinerjanya dalam mendeteksi clickbait judul berita dengan lebih akurat dan efektif.

Beberapa ciri suatu artikel tergolong clickbait atau bukan clickbait [10], yaitu: pertama, menggunakan judul yang menarik atau sensasional agar terkesan heboh, padahal isi beritanya biasa saja; ciri kedua, merupakan ciri yang cukup sering dijumpai pada artikel clickbait, judul yang tidak sesuai dengan isinya. Contohnya “Begini cara menghilangkan jerawat dalam sekejap”, ternyata pada saat diklik dan dibaca, artikel hanya membahas bahan-bahan yang baik untuk wajah tanpa ada penjelasan bagaimana cara menggunakannya. Ketiga, menggunakan kata-kata bombastis seperti frasa seperti ‘Wow!’, ‘Heboh!’ atau ‘Klik di sini!’ membuat pembaca terpancing untuk mengklik artikel. Keempat, mengundang minat segmen pembaca tertentu dengan menggunakan kata-kata golongan atau usia tertentu seperti (18+), (Khusus Dewasa) atau (Yang Merokok Jangan Baca!). Kelima, Judul yang dilebih-lebihkan, yakni dengan menggunakan kata-kata superlatif seperti ‘terkaya’, ‘tercantik’, ‘terkenal’, ‘terbaik’ dan ‘paling’. Keenam, menggunakan tanda baca yakni tanda kutip (“...”) yang bertujuan sebagai makna kiasan untuk memancing prasangka dan rasa penasaran pembaca yang ingin mengetahui makna asli dari kata-kata yang diberi tanda kutip tersebut. Yang terakhir yakni Elipsis. Penggunaan tanda baca elipsis (...) seperti “SBY Setuju jadi Cawapres, asal...”. Penggunaan tanda baca elipsis ini tidak selalu bersifat menjebak, akan tetapi lebih ke arah memancing rasa penasaran pembaca agar mendapatkan jawaban utuh dari judul ketika membaca artikelnya.

3. Hasil dan Pembahasan

3.1 Analisa dan Pemecahan Masalah

Kemudahan akses internet saat ini telah mempermudah proses berbagi berita. Berbeda dengan masa lalu yang terbatas pada media cetak dan elektronik, sekarang, penyampaian dan penyebaran berita menjadi lebih mudah dan praktis melalui internet. Dengan hadirnya internet, situs-situs berita bermunculan dan jumlahnya terus bertambah. Setiap hari, jumlah berita yang dibagikan sangatlah banyak, mulai dari yang bersifat tidak penting hingga berita yang sangat penting [14].

Dalam penulisan berita, perhatian terhadap judul sangat penting. Salah satu masalahnya adalah kecenderungan penulis atau penyedia berita online untuk membuat judul yang terlalu berlebihan atau provokatif, yang bertujuan untuk menarik perhatian pembaca. Bahkan, berita yang sebenarnya tidak terlalu penting pun dapat terlihat signifikan dengan penggunaan judul yang menarik [15]. Hal ini seringkali menyebabkan pembaca merasa tertipu ketika mereka membaca judul yang

berlebihan, yang tidak selaras dengan isi berita yang sebenarnya., Ketika membaca isinya, ternyata berita itu biasa saja atau bahkan tidak penting sama sekali.

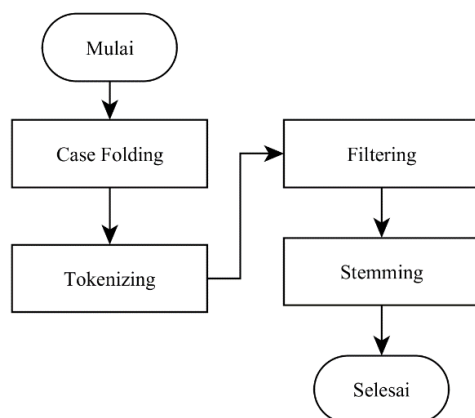
Proses klasifikasi ini, berita akan dibagi menjadi dua kelas atau kelompok, yaitu clickbait dan non-clickbait. Metode yang akan digunakan dalam penelitian ini adalah K-NN, di mana model akan mencari tetangga terdekat dari data uji berdasarkan sejumlah parameter k yang ditentukan. Diharapkan dengan menggunakan pendekatan ini, akan lebih mudah untuk mengidentifikasi judul-judul berita yang bersifat clickbait dan meminimalkan kesan menyesatkan bagi pembaca.

Dari hasil klasifikasi tersebut, akan teridentifikasi judul berita yang termasuk dalam kategori clickbait dan yang bukan clickbait. Judul-judul yang tergolong clickbait cenderung menggunakan gaya penulisan yang provokatif atau sensasional untuk menarik perhatian pembaca, sementara judul non-clickbait lebih fokus pada penyampaian informasi yang jelas dan objektif.

Evaluasi keakuratan K-NN dalam penelitian ini akan dilakukan dengan membandingkan prediksi klasifikasi yang diberikan oleh model dengan label sebenarnya dari data pengujian. Metrik evaluasi seperti akurasi, presisi, recall, dan F1-score akan digunakan untuk mengukur seberapa baik model dapat membedakan antara judul-judul clickbait dan non-clickbait [13]. Semakin tinggi nilai matriks tersebut, semakin baik kinerja model dalam melakukan klasifikasi.

3.2 Proses Text Processing

Text preprocessing seperti terlihat pada Gambar 3 merupakan tahap awal dalam pelatihan K-NN. Dalam text preprocessing sendiri terdapat tahap-tahap yang harus dijalani, yaitu case folding, tokenizing, filtering, dan stemming [1].



Gambar 3. Proses Text Processing

Proses case folding dilakukan dengan mengubah semua huruf dalam teks menjadi huruf kecil tanpa memperhatikan aturan penulisan huruf [9]. Meskipun kata awal dalam kalimat, nama orang, dan nama kota

yang biasanya ditulis dengan huruf kapital, pada proses case folding semuanya akan diubah menjadi huruf kecil. Tabel 2 merupakan contoh judul berita yang melalui tahap proses case folding.

Tabel 2. Sub Proses Case Folding

Judul Berita	Tahap
Diprotes, Ini Jawaban Nike Soal Produk Hijab	Kurasi Data
diprotes, ini jawaban nike soal produk hijab aksi 313	Case Folding

Proses tokenisasi berfungsi dengan memisahkan setiap kalimat dalam judul berita menjadi kata-kata yang mandiri dan tidak bergantung pada kata lainnya. Proses selanjutnya dilanjutkan dengan proses filtering dengan proses stopwords removal yakni kata-kata yang dianggap kurang penting seperti kata penghubung, kata tanya dan kata umum lainnya yang dianggap tidak memiliki makna [16]. Akan tetapi dalam penelitian ini akan ditambahkan daftar kata yang merupakan indikasi atau ciri dari judul berita clickbait “ini”, “begini”, kata tanya, dan beberapa kata lainnya yang termasuk kata unik clickbait, sehingga kata-kata tersebut tidak akan dihilangkan dalam proses stopwords removal. Tahap terakhir adalah stemming, cara kerja stemming adalah dengan mengubah setiap kata berimbuhan menjadi kata dasar [17], sedangkan kata yang sudah tak berimbuhan dalam artian sudah berbentuk kata dasar akan diabaikan. Tabel 3 adalah contoh judul berita yang sudah melalui proses stopwords removal, terdapat penghilangan beberapa kata, misalnya kata aksi dan 313.

Tabel 3. Sub Proses Stopword Removal

Judul Berita	Tahap
diprotes, ini jawaban nike soal produk hijab aksi 313	Case Folding
protes, ini nike soal produk hijab	Stopwords Removal

3.3 Implementasi Model K-NN

Pada proses ini terdiri dari tiga fase, fase pertama yakni pembobotan TF-IDF yang merupakan fitur seleksi dari model K-NN untuk mendeteksi clickbait. Pembobotan digunakan untuk memberikan bobot pada setiap kata yang sudah melalui tahap text preprocessing pada masing-masing data [1], baik data latih maupun data uji. Proses pembobotan menggunakan TF-IDF dengan tahap sebagai berikut. Pertama Term Frequency (TF), merupakan jumlah kemunculan suatu term atau kata pada setiap artikel. Kedua, Document Frequency (DF), merupakan jumlah data yang mengandung satu kata tertentu. Ketiga Inverse Document Frequency (IDF), merupakan pengurangan bobot suatu term yang tingkat kemunculannya tinggi di berbagai dokumen yang ada. IDF diperoleh dengan rumus $\log(n/DF)$, di mana n merupakan jumlah data [16]. Hasil perhitungan pembobotan setelah dilakukan perhitungan jarak (*Euclidean Distance*) disajikan pada Tabel 8.

Tahap berikutnya adalah perhitungan Jarak (Euclidean Distance), menghitung jarak setiap kata dalam satu judul terhadap judul berita lainnya. Hal ini diperlukan untuk mengetahui apakah hasil akhir klasifikasi sesuai dengan label yang diberikan pada tahap pengumpulan data. Perhitungan jarak pada penelitian ini akan menggunakan metode Euclidean Distance. Tahap yang terakhir adalah proses pencarian tetangga Terdekat (K-NN). Setelah jarak dari judul berita ke-1 terhadap setiap judul berita lainnya telah diperoleh, selanjutnya adalah mengurutkan hasil tersebut dari yang terkecil hingga terbesar. Pada Euclidean Distance semakin besar jarak, semakin rendah tingkat kemiripannya [12]. Begitu juga sebaliknya, semakin kecil jarak, semakin tinggi pula tingkat kemiripannya.

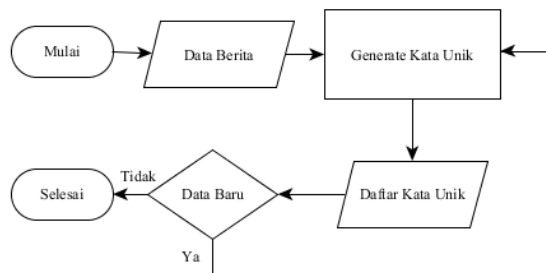
Tabel 4. Hasil Perhitungan K-NN

Jarak	Judul Berita ke-	Clickbait
3.3404126	8	Ya
3.6055513	3	Tidak
3.6274449	6	Ya
3.6726774	7	Tidak
3.7627591	2	Ya
4	4	Tidak
4.3588989	5	Tidak

Tabel 4 menunjukkan hasil perhitungan K-Nearest Neighbors (K-NN) untuk klasifikasi judul berita sebagai clickbait atau bukan. Hasil perhitungan ini digunakan untuk menentukan klasifikasi judul berita berdasarkan jarak terdekat ke data pelatihan yang telah dilabeli.

3.4 Kata Unik

Kata-kata unik berperan sebagai fitur dan indikator pendukung dalam mengidentifikasi judul Clickbait. Informasi ini diperoleh melalui proses pencarian dan perbandingan antara kata-kata yang terdapat dalam judul Clickbait namun tidak ditemukan dalam judul Bukan Clickbait [10]. Oleh karena itu, data ini menjadi elemen penting dalam membedakan dan mengklasifikasikan judul-judul Clickbait. Setiap kali terjadi perubahan dalam judul, label, penambahan, atau penghapusan data, penting untuk melakukan pencarian ulang terhadap kata-kata unik tersebut karena kemungkinan adanya perubahan dalam data kata unik.



Gambar 4. Flowchart Kata Unik

Gambar 4 merupakan *flowchart* atau alur yang menggambarkan bagaimana tahap pengumpulan data kata unik.

Tabel 5. Tabel Data Kata Unik

No.	Kata Unik	Jumlah
1	apa	16
2	kenapa	12
3	nih	11
4	alasan	11
5	wow	10
6	gara	10
7	heboh	10
8	siapa	7
9	kamu	6
10	alasannya	6
11	mengapa	6
12	komentar	6
13	lihat	6
14	disebut	5
15	tahu	5
16	kok	5
17	awas	5
18	inilah	5
19	begini	5
20	wah	4

Tabel 5 adalah hasil perolehan kata unik yang dikerjakan oleh program dalam penelitian yang telah dilakukan. Kata-kata seperti "apa", "kenapa", "nih", dan "alasan" merupakan kata-kata yang paling sering muncul, dengan jumlah kemunculan yang cukup signifikan. Sebaliknya, kata-kata seperti "siapa", "kamu", "alasannya", dan "mengapa" memiliki kemunculan yang lebih sedikit. Ini menunjukkan bahwa kata-kata pertanyaan dan kata-kata yang terkait dengan perasaan atau reaksi seperti "wow", "gara", dan "heboh" juga sering muncul. Kata "wah" dan "kok" tingkat kemunculannya jauh lebih sedikit jika dibandingkan kata "wow" dan "nih", walaupun kata-kata tersebut termasuk dalam kata reaksi. Hal yang hampir sama, terjadi juga pada kata-kata yang termasuk dalam kelompok kata tanya, seperti kata "apa", "kenapa", "siapa" dan kata "mengapa".

3.5 Pengujian Algoritma

Pada tahap pengujian algoritma, hasil akurasi terbaik didapat pada jumlah tetangga atau $k = 11$. Tabel 6 adalah persentase evaluasi hasil pengujian algoritma K-NN terhadap 200 data uji.

Dari hasil pengujian menggunakan confusion matrix yang disajikan pada tabel 6, terlihat bahwa variasi jumlah tetangga (K) memiliki pengaruh yang signifikan terhadap kinerja sistem. Dalam tabel tersebut, terdapat nilai true positive (TP), true negative (TN), false positive (FP), dan false negative (FN) untuk setiap nilai K yang diuji. Dari nilai-nilai tersebut, dapat dihitung metrik evaluasi seperti Precision, Recall, dan F-Measure. Dapat dilihat bahwa dengan peningkatan jumlah tetangga, kinerja sistem cenderung meningkat, terutama dalam hal akurasi dan F-Measure [17]. Hasil terbaik tercapai pada nilai $K=11$, dengan akurasi

sebesar 80.5%, Precision sebesar 85%, Recall sebesar 81%, dan F-Measure sebesar 80%. Hal ini menunjukkan bahwa model dengan jumlah tetangga 11

memberikan keseimbangan yang baik antara presisi dan recall dalam mengklasifikasikan judul-judul clickbait.

Tabel 6. Evaluasi Pengujian Algoritma

K	TP	TN	FP	FN	Confusion Matrix	Precision	Recall	F-Measure
3	42	88	3	67	65%	77%	65%	62%
5	57	88	3	52	72.5%	80%	72%	71%
7	64	87	4	45	75.5%	81%	76%	75%
9	66	86	5	43	76%	81%	76%	76%
11	72	89	2	37	80.5%	85%	81%	80%
13	73	85	6	36	79%	82%	79%	79%
15	68	85	6	41	76.5%	81%	77%	76%

3.6 Pengujian Menggunakan Data Baru

Setelah mencapai tingkat akurasi minimal 80% dengan pengujian menggunakan data uji, langkah selanjutnya adalah melakukan pengujian dengan menggunakan data baru [17]. Pengujian dengan data baru dilakukan dengan mengakses halaman utama, kemudian memasukkan judul berita yang akan diklasifikasikan. Pada tahap ini, dilakukan pengujian dengan menggunakan 20 judul berita baru, dan hasilnya disajikan dalam Tabel 7.

Tabel 7. Evaluasi Pengujian Data Baru

TP	TN	FP	FN	Confusion Matrix	Precision	Recall	F-Measure
9	9	1	1	90%	90%	90%	90%

Hasil pengujian training yang diteruskan dengan pengujian menggunakan data baru menunjukkan hasil yang lebih signifikan. Dari tabel yang diberikan, terlihat bahwa model berhasil mengklasifikasikan data baru dengan tingkat akurasi yang tinggi. Nilai TP, TN, FP, dan FN yang menggambarkan kinerja model dalam mengklasifikasikan data baru. Dengan menggunakan metrik evaluasi, seperti Precision, Recall, dan F-Measure, dapat disimpulkan bahwa model memiliki performa yang sangat baik dalam mengenali judul-judul clickbait pada data baru. Dengan tingkat akurasi mencapai 90%, serta nilai Precision, Recall, dan F-Measure yang sama-sama tinggi sebesar 90%, menunjukkan bahwa model dapat mengidentifikasi clickbait dengan baik tanpa banyak mengorbankan presisi atau recall. Hal ini menunjukkan bahwa model yang telah dilatih dengan data sebelumnya mampu beradaptasi dengan baik dalam mengklasifikasikan data baru.

4. Kesimpulan

Setelah melalui tahapan analisis, perancangan, implementasi, pelatihan, pengujian, dan evaluasi terhadap sistem pengklasifikasi judul berita berbahasa Indonesia menggunakan metode k-Nearest Neighbor, dapat disimpulkan bahwa aplikasi atau sistem ini mampu mengklasifikasikan judul berita berbahasa

Indonesia ke dalam kelas Clickbait atau Bukan Clickbait. Faktor-faktor yang mempengaruhi kinerja sistem termasuk jumlah tetangga (k), variasi data, dan fitur pendukung yang digunakan di dalamnya. Konfigurasi terbaik algoritma K-NN untuk penelitian ini adalah $k = 11$ dengan data latih 800 artikel dan data uji 200 artikel, yang menghasilkan akurasi sebesar 80.5%, Presisi 85%, Recall 81%, dan F-Measure 80%. Selain itu, beberapa kata unik yang sering muncul pada judul Clickbait dalam penelitian ini antara lain "apa", "kenapa", "nih", "alasan", "wow", dan beberapa kata lainnya yang terdapat pada daftar kata unik. Selanjutnya, dilakukan pengujian menggunakan 20 judul baru yang belum memiliki kelas dengan hasil akurasi sebesar 90%.

Daftar Rujukan

- [1] N. A. Zuhroh and N. A. Rakhmawati, "Clickbait detection: A literature review of the methods used," *Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 6, no. 1, pp. 1-10, 2020.
- [2] M. N. Fakhruzzaman, S. Z. Jannah and R. A. Ningrum, "Flagging clickbait in Indonesian online news websites using fine-tuned transformers," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 3, 2023.
- [3] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Systems with Applications*, vol. 153, 2020.
- [4] J. Sirusstara, N. Alexander, A. Alfariy, S. Achmad and R. Sutoyo, "Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa)," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, Ipoh, Malaysia, 2022.
- [5] Y. Fahrimal, A. Husna, F. Islami and J. Johan, "MEDIA DAN PANDEMI: FRAME TENTANG PANDEMI COVID-19 DALAM," *JURNAL STUDI KOMUNIKASI DAN MEDIA*, vol. 24, no. 2, pp. 169-186, 2020.
- [6] J. Zheng, K. Yu and X. Wu, "A deep model based on Lure and Similarity for Adaptive Clickbait Detection," *Knowledge-Based Systems*, vol. 214, 2021.
- [7] M. Dong, L. Yao, X. Wang, B. Benatallah and C. Huang, "Similarity-Aware Deep Attentive Model for Clickbait Detection," in *Pacific-Asia Conference on Knowledge*

- Discovery and Data Mining PAKDD 2019: Advances in Knowledge Discovery and Data Mining*, Macau, China, 2019.
- [8] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," *IEEE Access*, vol. 8, pp. 61726-61740, 2020.
- [9] A. William and Y. Sari, "CLICK-ID: A novel dataset for Indonesian clickbait headlines," *Data in Brief*, vol. 32, p. 106231, 2020.
- [10] B. Naeem, A. Khan, M. O. Beg and H. Mujtaba , "A deep learning framework for clickbait detection on social area network using natural language cuesA deep learning framework for clickbait detection on social area network using natural language cues," *Journal of Computational Social Science*, vol. 3, p. 231–243 , 2020.
- [11] N. Kaothanthong, S. Kongyoung and T. Theeramunkong, "Headline2Vec: A CNN-based Feature for Thai Clickbait Headlines Classification," *INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY*, vol. 5, no. 1, 2021.
- [12] T. Mladenova and I. Valova, "Analysis of the KNN Classifier Distance Metrics for Bulgarian Fake News Detection," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Turkey, 2021.
- [13] I. Ahmad, M. A. Alqarni, A. A. Almazroi and A. Tariq, "Experimental Evaluation of Clickbait Detection Using Machine Learning Models," *Intelligent Automation & Soft Computing* , vol. 26, no. 6, pp. 1335-1344, 2020.
- [14] W. Wang, F. Feng, X. He, H. Zhang and T.-S. Chua, "Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue," in *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Canada, 2021.
- [15] V. Kaushal and K. Vemuri, "Clickbait—Trust and Credibility of Digital News," *IEEE Transactions on Technology and Society* , vol. 2, no. 3, pp. 146-154, 2021.
- [16] M. Liebenlito, A. A. Yesinta and M. I. S. Musti, "Deteksi Clickbait pada Judul Berita Online Berbahasa Indonesia Menggunakan FastText," *Journal of Applied Computer Science and Technology (JACOST)* , vol. 5, no. 1, 2024.
- [17] D. K. Dixit, A. Bhagat and D. Dangi, "An accurate fake news detection approach based on a Levy flight honey badger optimized convolutional neural network model," *Concurrency and Computation Practice and Experience*, vol. 35, no. 1, 2023.

Lampiran

Tabel 8. Pembobotan Kata

TERM	Jarak Judul ke-1 Judul Berita						
	2	3	4	5	6	7	8
begini	1	1	1	1	1	1	1
kondisi	1	1	1	1	1	1	1
jalan	1	1	1	1	1	1	1
sekitar	1	1	1	1	1	1	1
masjid	1	1	1	1	1	1	1
istiqlal	1	1	1	1	1	1	1
aksi	1	1	1	1	1	1	1
313	1	1	1	1	1	1	1
ini	0.1583563	0	0	0	0.1583563	0	0.1583563
cassandra	1	0	0	0	0	0	0
lee	1	0	0	0	0	0	0
susah	1	0	0	0	0	0	0
paling	1	0	0	0	0	0	0
randy	1	0	0	0	0	0	0
martin	1	0	0	0	0	0	0
ahok	0	1	0	0	0	0	0
lift	0	1	0	0	0	0	0
bunyi	0	1	0	0	0	0	0
paksa	0	1	0	0	0	0	0
masuk	0	1	0	0	0	0	0
sandiaga	0	0	1	0	0	0	0
uno	0	0	1	0	0	0	0
pan	0	0	1	0	0	0	0
tulang	0	0	1	0	0	0	0
punggung	0	0	1	0	0	0	0
program	0	0	1	0	0	0	0
ok	0	0	1	0	0	0	0
oce	0	0	1	0	0	0	0
kevin	0	0	0	1	0	0	0

DOI: <https://doi.org/10.38204/tematik.v11i1.1872>

Lisensi: Creative Commons Attribution 4.0 International (CC BY 4.0)

TERM	Jarak Judul ke-1						
	Judul Berita						
	2	3	4	5	6	7	8
sanjaya	0	0	0	1	0	0	0
sukamuljo	0	0	0	1	0	0	0
marcus	0	0	0	1	0	0	0
fernaldi	0	0	0	1	0	0	0
gideon	0	0	0	1	0	0	0
juara	0	0	0	1	0	0	0
ganda	0	0	0	1	0	0	0
all	0	0	0	1	0	0	0
england	0	0	0	1	0	0	0
2017	0	0	0	1	0	0	0
protes	0	0	0	0	1	0	0
nike	0	0	0	0	1	0	0
soal	0	0	0	0	1	0	0
produk	0	0	0	0	1	0	0
hijab	0	0	0	0	1	0	0
ridwan	0	0	0	0	0	1	0
kamil	0	0	0	0	0	1	0
usung	0	0	0	0	0	1	0
nasdem	0	0	0	0	0	1	0
cagub	0	0	0	0	0	1	0
jabar	0	0	0	0	0	0.4885591	0
apa	0	0	0	0	0	0	1
minum	0	0	0	0	0	0	1
mudah	0	0	0	0	0	0	1
angka	0	0	0	0	0	0	0
anggur	0	0	0	0	0	0	0
hati	0	0	0	0	0	0	0
anak	0	0	0	0	0	0	0
pedas	0	0	0	0	0	0	0
ussy	0	0	0	0	0	0	0
sulistiawaty	0	0	0	0	0	0	0
JUMLAH	14.158356	13	16	19	13.158356	13.488559	11.158356
HASIL AKAR	3.7627591	3.6055513	4	4.3588989	3.6274449	3.6726774	3.3404126