Analisis Sentimen terhadap Opini Publik Mengenai *Childfree* dalam Pernikahan pada Twitter Menggunakan *K-Nearest Neighbor* (K-NN)

Aries Dwi Indriyanti[1], Putri Nur Fadhilah[2]

[1,2] Program Studi Teknik Informatika, Universitas Negeri Surabaya
[1]ariesdwi@unesa.ac.id
[2]putri.19090@mhs.unesa.ac.id

***Abstrak*—Pada pernikahan biasanya yang paling di nanti ialah memiliki seorang keturunan, sebagai tanda kesempurnaan dan menambah kebahagiaan didalam sebuah pernikahan. Namun menurut perkembangan terkini, banyak pasangan pernikahan memilih untuk menjalani kehidupan yang bahagia tanpa memiliki anak atau disebut dengan *childfree*. Tetapi masyarakat masih banyak yang memperdebatkan hal ini, salah satunya di twitter. Maka dari itu diperlukan analisis sentimen guna melakukan prediksi opini publik yang cenderung positif, netral, atau negatif. *Preprocessing* teks yang digunakan adalah *cleaning*, normalisasi, *case folding*, *filtering*, *stemming*, dan *tokenizing*. Dimana tujuan dari normalisasi adalah mengganti kata-kata yang disingkat dan teknik pembobotan katanya menerapkan TF-IDF. Data pada penelitian ini berjumlah 1100 tweet. Dengan menerapkan algoritma *K-Nearest Neighbor* dan diuji untuk menghasilkan sistem yang paling akurat. Hasil dari pelabelan untuk sentimen positif berjumlah 636, sentimen netral berjumlah 285, dan 285 untuk sentimen negatif. Data di-*split* dengan perbandingan 85% untuk data *train* dan 15% untuk data *test*. Klasifikasi K-NN dengan nilai k = 8 menghasilkan akurasi sebesar 70%, *precision* 70% dan *recall* 70%.**

***Kata Kunci*—Sentimen Analisis, K-Nearest Neighbor (K-NN), Childfree**

I. PENDAHULUAN

Pada pernikahan biasanya yang paling di nanti ialah memiliki seorang keturunan, sebagai tanda kesempurnaan dan menambah kebahagiaan didalam sebuah pernikahan. Namun menurut perkembangan terkini, banyak pasangan pernikahan memilih untuk menjalani kehidupan yang bahagia tanpa memiliki anak *childfree*. Tetapi masyarakat masih banyak yang memperdebatkan hal ini. Komentar yang didorong oleh perasaan penulisnya akan dikategorikan untuk menentukan masuk dalam kelompok sentimen yang mana. Untuk itu dengan melihat permasalahan yang ada, diperlukan penelitian untuk melakukan analisis sentimen. Analisis sentimen pada penelitian ini dilakukan dengan menerapkan metode *K-Nearest Neighbor* (K-NN). K-NN mempunyai keuntungan karena mudah dipahami dan memberikan hasil prediksi yang akurat. Metode ini mampu melakukan klasifikasi data secara tepat dengan memilih jumlah tetangga terdekat yang sesuai dengan nilai K yang dipilih.

Penelitian terkait yang pernah dilakukan yaitu implementasi K-NN untuk sentimen analisis publik dalam pembelajaran daring. Penelitian tersebut diperoleh hasil yang paling akurat terjadi pada nilai k = 10 yang menghasilkan akurasi sebesar 84,65%, *f-score* sebesar 87%, *precision* sebesar 87%, dan *recall* sebesar 86%, serta tingkat *error* sebesar 0,12% [1]. Penelitian lainnya adalah penggunaan K-NN untuk analisis sentimen pada akun Twitter PT PLN (Persero) kualitas yang menghasilkan akurasi sebesar 87,41% [2]. Penelitian selanjutnya yaitu analisis sentimen capres Indonesia 2019 di Twitter menggunakan metode K-NN yang menghasilkan tingkat akurasi sebesar 83,33% [3]. Adapun penelitian metode K-NN dan pembobotan *term* untuk analisis sentimen Twitter pada layanan siakad Universitas Brawijaya yang menghasilkan akurasi saat menggunakan k=3 sebesar 86% [4]. Analisis sentimen lainnya menggunakan algoritma K-NN pada objek wisata Dufan. Penelitian ini menunjukkan hasil pada saat nilai k = 7 mendapatkan tingkat akurasi tertinggi 77,01%, nilai kurva AUC 0,894, presisi 92,38%, dan *recall* 61,56% [5].

Dari temuan di atas, penelitian ini menggunakan *K-Nearest Neighbor* (KNN), metode sederhana yang sering digunakan untuk klasifikasi teks dan data, bersama dengan pembobotan TF-IDF. Data pada penelitian ini terdiri dari tweet pendapat masyarakat mengenai *childfree* dalam pernikahan yang diambil dari media sosial Twitter. Proses *preprocessing* yang dilakukan pada data tweet sebelum digunakan termasuk membersihkan, normalisasi, *case folding*, *filtering*, *stemming*, dan *tokenizing*. Tujuan normalisasi adalah memperbaiki kata yang disingkat. 15% data uji dan 85% data latih masing-masing. Sebagai evaluasi hasil dari metode ini digunakan matriks yang terdiri dari akurasi, *precision*, dan *recall*.

II. METODE PENELITIAN

A. Pengumpulan Data

Data dikumpulkan dengan menggunakan Octoparse v8.6, proses pengumpulan data hanya membahas topik atau pendapat tentang *childfree* pada media sosial Twitter. Data yang dikumpulkan dari Twitter mencakup 1100 tweet dengan kata kunci "*childfree*" dalam bahasa Indonesia. Tabel berikut menunjukkan jumlah data dari setiap opininya.

TABEL I
JUMLAH DATA TWEET

Opini	Jumlah Data
Positif	636
Netral	179

Negatif	285
Total	**1100**

B. Preprocessing Data

Preprocessing adalah sebuah tahap yang bermanfaat dalam teks *mining*. Tujuan dari *preprocessing* adalah untuk menjamin data yang digunakan pada analisis adalah data yang berkualitas tinggi dan representatif. Jika data yang digunakan dalam proses penemuan pengetahuan memiliki kualitas rendah, maka hasil pengetahuan yang diperoleh juga akan rendah [6]**.** Pada penelitian ini tahapan *preprocessing* yang dilakukan meliputi:

1) Cleaning

Cleaning adalah tahapan yang melibatkan penghapusan baris atau kolom yang tidak memiliki nilai atau memiliki nilai yang hilang.

2) Case Folding

Merupakan proses mengganti seluruh teks menjadi huruf kecil. Ini dilakukan dalam *preprocessing* data teks untuk menghilangkan perbedaan antara huruf besar dan huruf kecil, yang bisa mempengaruhi hasil analisis teks. Proses ini bertujuan untuk memastikan bahwa teks yang sama diterima sebagai teks yang sama, meskipun ada perbedaan dalam penulisan huruf besar dan huruf kecil.

3) Tokenizing

Tokenizing adalah tahap pemotongan teks per kata sesuai dengan kamus data yang ditetapkan. *Tokenizing* dilakukan untuk mendapatkan kata-kata yang bernilai dan mempermudah dalam menemukan frekuensi data dalam *corpus*.

4) Normalisasi

Normalisasi kata adalah tahap pembersihan dan memperbaiki kata singkatan menjadi kata yang bermakna sama, berdasarkan Kamus Besar Bahasa Indonesia (KBBI). Hal ini dilakukan untuk membuat informasi lebih mudah diproses dan memudahkan dalam tahap analisis data selanjutnya.

5) Filtering

Adalah tahap pembersihan kata tidak penting pada suatu dokumen teks dengan cara membandingkannya dengan daftar *Stopword (Stoplist)* yang tersedia. Pada penelitian ini, tahap *filtering* dilakukan menggunakan algoritma *stopword removal*.

6) Stemming

Stemming merupakan proses dalam *preprocessing* teks yang bertujuan untuk menghilangkan suffiks (akhiran) pada suatu kata untuk memperoleh kata dasar (*root word*). *Stemming* sangat penting dalam aplikasi teks mining, seperti klasifikasi dokumen, ekstraksi informasi, dan pencocokan pola. Ini membantu menghilangkan variasi dari kata yang sama sehingga mempermudah dalam proses analisis teks.

C. K-Nearest Neighbor (K-NN)

Algoritma K-NN merupakan algoritma sederhana yang sering digunakan untuk klasifikasi teks dan data [7]. K-NN digunakan untuk melakukan prediksi terhadap data sesuai dengan fitur dari kelas yang serupa. Dalam K-NN, kelas suatu objek baru di dalam data *training* diklasifikasikan berdasarkan mayoritas kelas dari K tetangga terdekatnya.

Algoritma K-NN mengidentifikasi objek terdekat dengan data yang diklasifikasikan. Lalu, data dikelompokkan ke dalam suatu kelas berdasarkan kemungkinan yang paling tinggi [8]. Untuk mengetahui jarak antara data pertanyaan dan data pembelajaran, rumus jarak geometris digunakan untuk menghitung jarak antara titik data pertanyaan dan semua titik data pembelajaran. Kemudian, dihitung sejumlah k buah terdekat dengan data pertanyaan. Diproyeksikan bahwa titik yang baru diklasifikasikan akan termasuk dalam titik yang diklasifikasikan terbanyak. metode *Euclidean Distance* digunakan dalam melakukan perhitungan jarak antar data dengan persamaan (1).

$$dist = \sum_{i=1}^{p} \sqrt{(x2 - x1)^2} \quad (1)$$

Keterangan:

x1 : Data *train*

x2 : Data *test*

dist : Jarak

i : Indeks penjumlahan

p : Jumlah atribut

D. Pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF)

Pembobotan TF-IDF adalah algoritma pengukuran bobot kata pada dokumen atau kumpulan dokumen untuk menentukan tingkat penting tidaknya kata tersebut pada konteks dokumen. Pembobotan ini digunakan untuk memudahkan pemrosesan teks dalam analisis teks dan pengambilan informasi.

Pembobotan TF-IDF mengukur banyak munculnya kata dalam suatu dokumen lalu mengkombinasikannya dengan jumlah kemunculan kata tersebut dalam kumpulan dokumen. Bobot TF-IDF semakin tinggi jika kata tersebut sering ada pada dokumen tertentu, namun jarang ada dalam dokumen lain dalam kumpulan dokumen. Pembobotan ini dapat membantu mengidentifikasi kata-kata kunci dalam dokumen dan mempermudah pencarian dan klasifikasi dokumen.

Dalam pembobotan TF-IDF, *Term Frequency* (TF) adalah frekuensi kata pada dokumen, dan *Inverse Document Frequency* (IDF) adalah logaritma dari jumlah dokumen dalam kumpulan dokumen dibagi dengan jumlah dokumen yang mengandung kata tersebut. Bobot TF-IDF didapatkan dengan melakukan perkalian antara TF dan IDF. TF-IDF

digunakan untuk memungkinkan analisis dengan algoritma K-NN. Berikut adalah rumus TF-IDF yang dintunjukkan oleh persamaan (2) dan (3).

$$idf = \log \frac{N}{df} \quad (2)$$

$$w(k,d) = tf(k,d) * idf \quad (3)$$

$$w(k,d) = tf(k,d) * \log \frac{N}{df}$$

Keterangan :

tf(k,d) : jumlah kata yang muncul dalam dokumen

W(k,d) : bobot kata pada dokumen

df : jumlah dokumen mengandung *term*

N : jumlah total dokumen di *database*

III. Hasil dan Pembahasan

Tujuan dari penelitian ini adalah mendapatkan tingkat akurasi dalam melakukan analisis sentimen terhadap opini publik tentang *childfree* dalam pernikahan di media sosial twitter dengan menerapkan algoritma K-NN.

A. *Preprocessing Data*

Cleaning, tokenizing, case folding, normalisasi*, filtering,* dan *stemming* adalah semua prosedur *preprocessing data.* Tabel berikut adalah contoh hasil *preprocessing data*.

Tabel III
Preprocessing Data

Tahapan	Hasil
Data Tweet Asli	Makanya org2 kaya zaman skrg makin marak punya satu anak. Bahkan ada yg pilih childfree. Mereka lbh aman ketika ekonomi tiba2 memburuk.
Cleaning	Makanya org kaya zaman skrg makin marak punya satu anak Bahkan ada yg pilih childfree Mereka lbh aman ketika ekonomi tiba memburuk
Case Folding	makanya org kaya zaman skrg makin marak punya satu anak bahkan ada yg pilih childfree mereka lbh aman ketika ekonomi tiba memburuk
Tokenizing	[makanya, org, kaya, zaman, skrg, makin, marak, punya, satu, anak, bahkan, ada, yg, pilih, childfree, mereka, lbh, aman, ketika, ekonomi, tiba, memburuk]
Normalisasi	[makanya, orang, kaya, zaman, sekarang, makin, marak, punya, satu, anak, bahkan, ada, yang, pilih, childfree, mereka, lebih, aman, ketika, ekonomi, tiba, memburuk]
Filtering	[makanya, orang, kaya, zaman, sekarang, makin, marak, punya, satu, anak, bahkan, pilih, childfree, lebih, aman, ekonomi, tiba, memburuk]
Stemming	[makanya, orang, kaya, zaman, sekarang, makin, marak, punya, satu, anak, bahkan, pilih, childfree, lebih, aman, ekonomi, tiba, memburuk]
Data Tweet Akhir	makanya orang kaya zaman sekarang makin marak punya satu anak bahkan pilih childfree lebih aman ekonomi tiba memburuk

B. *Visualisasi Data*

Setelah dilakukan *preprocessing data*, berikut adalah visualisasi data berupa *wordcloud* untuk keseluruhan data, opini positif, opini netral, dan opini negatif.

Gbr. 1 *Wordcloud* Seluruh Data

Gbr. 2 *Wordcloud* Opini Positif

Gbr. 3 *Wordcloud* Opini Netral

Gbr. 4 *Wordcloud* Opini Negatif

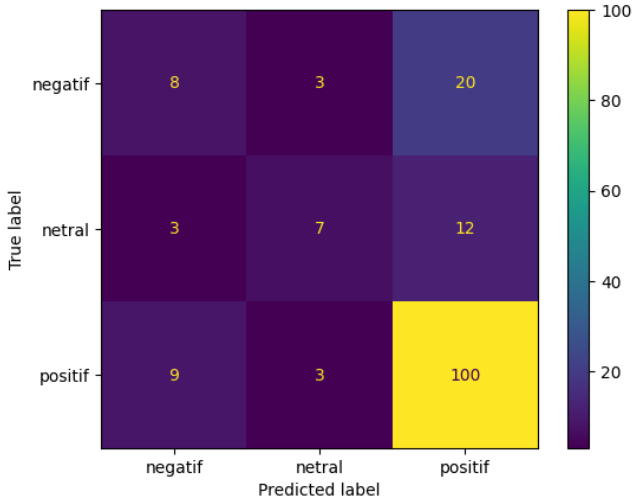
C. Pelatihan dan Pengujian

Data hasil *preprocessing* kemudian dipecah menjadi data *train* dan *test*. Pada penelitian ini digunakan perbandingan 85:15 untuk data *train* dan *test*. Akibatnya data *train* berjumlah 935 dan jumlah data *test* adalah 165.

Digunakan nilai k = 8 untuk klasifikasi dengan menerapkan K-NN pada penelitian ini. Dengan data *train* dan data *test* yang telah dihitung bobotnya dengan menerapkan pembobotan TF-IDF.

D. Evaluasi Hasil

Dari hasil pengujian dilakukan evaluasi pada sistem. Pada penelitian ini metrik evaluasi yang digunakan meliputi akurasi, *precision*, dan *recall*. FiTabel berikut menunjukkan evaluasi dari sistem yang telah dibangun.



Gbr. 5 *Confusion Matrix*

TABEL IIIII
EVALUASI HASIL

Metrik Evaluasi	Nilai (%)
Akurasi	**70**
Precision	70
Recall	70

IV. PENUTUP

A. Kesimpulan

Analisis sentimen terhadap opini publik mengenai *childfree* dalam pernikahan di sosial media twitter dengan menggunakan K-NN dengan data tweet keseluruhan berjumlah 1100 data yang terdiri dari 636 data opini positif, 179 data opini netral, dan 285 data opini negatif, serta dilakukan pembagian untuk data latih dan uji dengan rasio masing-masing 85:15 menghasilkan nilai-nilai metrik evaluasi saat nilai k=8 yaitu akurasi sebesar 70%, *precision* sebesar 70%, dan *recall* sebesar 70%.

B. Saran

Agar dapat menghasilkan nilai metrik evaluasi yang lebih baik disarankan untuk lebih memperbanyak jumlah data sehingga menambah variasi data serta menyetarakan jumlah data setiap kelasnya. Tahapan *preprocessing* juga sangat mempengaruhi kualitas datanya sehingga disarankan untuk menggunakan tahapan maupun kamus yang lebih lengkap sehingga dapat meningkatkan kinerja sistem pada penelitian selanjutnya.

REFERENSI

[1] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 2, p. 121, 2021, doi: 10.22146/ijccs.65176.

[2] R. Damarta, A. Hidayat, and A. S. Abdullah, "The application of k-nearest neighbors classifier for sentiment analysis of PT PLN (Persero) twitter account service quality," *Journal of Physics: Conference Series*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012002.

[3] A. Z. Malik, E. Utami, and S. Raharjo, "Analisis Sentiment Twitter Terhadap Capres Indonesia 2019 dengan Metode K-NN," *Jurnal INFORMA Politeknik Indonesia Surakarta*, vol. 5, no. 2, pp. 1–7, 2019.

[4] L. R. Dharmawan, I. Arwani, and D. E. Ratnawati, "Analisis Sentimen pada Sosial Media Twitter Terhadap Layanan Sistem Informasi Akademik Mahasiswa Universitas Brawijaya dengan Metode K-Nearest Neighbor," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 3, pp. 959–965, 2020, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/7099

[5] R. Sari, "Analisis Sentimen Pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma K-Nearest Neighbor (K-Nn)," *EVOLUSI : Jurnal Sains dan Manajemen*, vol. 8, no. 1, pp. 10–17, 2020, doi: 10.31294/evolusi.v8i1.7371.

[6] J. Han, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[7] R. H. Satrio and M. A. Fauzi, "Klasifikasi Tweets Pada Twitter Menggunakan Metode K-Nearest Neighbour (K-NN) Dengan Pembobotan TF-IDF," vol. 3, no. 8, pp. 2548–964, 2019, [Online]. Available: http://j-ptiik.ub.ac.id

[8] A. Hanafi, A. Adiwijaya, and W. Astuti, "Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa

Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 9, no. 3, pp. 357–364, 2020, doi: 10.32736/sisfokom.v9i3.980.