# Pembentukan Dataset Token Sentimen Berdasarkan Akun Instagram Brand Elektronik Menggunakan K-Nearest Neighbors

**Kristian Adi Nugraha**
Program Studi Informatika, Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana
Jl. Dr. Wahidin Sudirohusodo 5-25, Yogyakarta 55224, Daerah Istimewa Yogyakarta, Indonesia
Email: adinugraha@ti.ukdw.ac.id

***Abstract. Generating Sentiment Token Dataset Based on Electronics Brand Instagram Account using K-Nearest Neighbors.*** *Instagram is currently one of the most popular social media platforms for businesses and brand owners to promote their products. Because Instagram is a two-way communication platform, people can respond to any promotional content posted on Instagram. People's reactions come in a variety of form, and frequently include both positive and negative sentiment. This study aims to identify the words used in one type of sentiment, then use the K-NN approach to construct a token dataset by summarizing the phrases in many labels according to the sentiment type. The total accuracy value of the dataset for K = 1 is 33.38% (positive), 59.96% (negative), and 56.60% (neutral) based on the results of the tests performed.*
***Keywords****: sentiment analysis, K-Nearest Neighbors, dataset, Instagram*

***Abstrak.*** *Instagram saat ini menjadi salah satu media sosial yang banyak digunakan oleh perusahaan atau pemilik brand untuk melakukan promosi terhadap produk-produk yang dimilikinya. Karena bersifat dua arah, masyarakat dapat memberikan respon terhadap aktivitas promosi yang dilakukan oleh sebuah perusahaan melalui Instagram. Respon dari masyarakat memiliki varian yang beragam dan seringkali mengandung unsur sentimen baik positif maupun negatif. Penelitian ini mencoba untuk mengidentifikasi kata-kata yang digunakan dalam satu jenis sentimen, kemudian membuat dataset token dengan cara merangkum kata-kata tersebut dalam beberapa label sesuai jenis sentimen masing-masing menggunakan metode K-NN. Berdasarkan hasil pengujian yang dilakukan, didapatkan nilai akurasi dari dataset sebesar 33.38% (positif), 59.96% (negatif), dan 56.60% (netral) untuk K = 1.*
***Kata Kunci:*** *analisis sentimen, K-Nearest Neighbors, dataset, Instagram*

## 1. Pendahuluan

Instagram merupakan salah satu media sosial yang banyak digunakan oleh masyarakat seluruh dunia, sebagaimana pada akhir tahun 2020 Instagram tercatat memiliki 1,16 miliar pengguna aktif [1] yang didominasi oleh pengguna berusia 25-34 tahun dengan persentase sebesar 31,5% [2]. Karena memiliki jumlah pengguna aktif yang cukup banyak, maka terdapat banyak perusahaan, *vendor*, atau *brand* yang mempromosikan produk-produk yang dimilikinya melalui *platform* Instagram. Selain menggunakan promosi dalam bentuk iklan berbayar, biasanya akun-akun *brand* tersebut juga secara rutin mempublikasikan konten baru pada akun Instagram-nya dalam bentuk *post* maupun Instagram *story*. Berbeda dengan promosi yang dilakukan melalui media televisi, radio, atau papan iklan yang ditujukan untuk seluruh masyarakat, promosi melalui media *internet* (salah satunya Instagram) memiliki kelebihan yaitu iklan hanya ditujukan kepada pengguna yang memiliki karakteristik sesuai dengan iklan yang dipublikasikan. Dengan demikian, promosi yang dilakukan melalui Instagram dapat dikatakan efektif karena mampu menjangkau banyak pengguna secara tepat.

Di dalam akun Instagram milik sebuah *brand*, khususnya untuk kategori *brand* elektronik, setiap *post* dapat dikategorikan ke dalam sebuah tipe tertentu bergantung pada konten yang terdapat di dalam *post* tersebut. Setelah akun tersebut mempublikasikan sebuah *post*, pengguna dapat memberikan respon terhadap *post* tersebut melalui komentar secara tertulis. Komentar-komentar yang dituliskan oleh pengguna memiliki maksud dan arti yang bermacam-macam. Sebagian dari komentar-komentar tersebut ada yang sekedar bertanya terkait produk pada

*post* tersebut. Namun, selain itu tidak sedikit pengguna yang memberi komentar bernada positif maupun negatif terkait produk tersebut yang mana hal ini disebut sebagai sentimen. Respon sentimen yang diberikan oleh pengguna tentunya bergantung pada isi dari *post* yang dipublikasikan oleh akun tersebut. Oleh sebab itu, dapat disimpulkan bahwa konten dari sebuah *post* dapat berpengaruh terhadap respon yang diberikan oleh masyarakat.

Salah satu kebutuhan dalam melakukan analisis sentimen adalah tersedianya *dataset* yang lengkap dan akurat karena luaran sentimen yang dihasilkan bergantung sepenuhnya kepada *dataset* yang digunakan. Permasalahan yang seringkali muncul terkait *dataset* adalah tidak tersedianya *dataset* yang tepat sehingga akurasi luaran yang dihasilkan tidak dapat memberikan hasil yang maksimal [3]. *Dataset* dalam bentuk pustaka yang disediakan oleh pihak ketiga biasanya memiliki bentuk yang terlalu umum, sementara permasalahan yang dihadapi oleh setiap orang berbeda-beda serta membutuhkan *dataset* yang spesifik [4]. Penelitian ini bertujuan untuk menghasilkan *dataset* sentimen dalam bentuk token-token kata yang menyatakan sebuah ekspresi sentimen. Pembentukan *dataset* dilakukan secara otomatis menggunakan metode *K-Nearest Neighbors* (*K-NN*) dengan objek pengujian berupa akun Instagram untuk *brand* atau *vendor* elektronik di Indonesia. *K-NN* merupakan salah satu metode kecerdasan buatan yang dapat digunakan untuk melakukan pengenalan atau pengelompokan data [5]. Data-data komentar dari beberapa akun Instagram *brand* elektronik akan dikumpulkan menjadi satu setelah dilakukan pra-pemrosesan sebelumnya. Kemudian, setiap data komentar akan diproses menggunakan metode *K-NN* untuk menentukan jenis sentimen dari komentar tersebut. Dari kelompok-kelompok sentimen yang berhasil terbentuk, penulis akan melakukan filterisasi terhadap token-token yang secara spesifik hanya terdapat pada satu kelompok sentimen saja. Kemudian, token tersebut akan dijadikan sebagai salah satu data pada *dataset*. Luaran dari penelitian ini diharapkan dapat memberikan kontribusi untuk penelitian-penelitian lain yang membutuhkan *dataset* secara spesifik.

## 2. Tinjauan Pustaka

Pemrosesan teks merupakan salah satu bidang ilmu komputer yang berisi berbagai proses terkait pengolahan data dalam bentuk teks atau dokumen menjadi bentuk lain dengan tujuan untuk memperoleh informasi baru yang lebih bermanfaat [6]. Salah satu sub-bidang pemrosesan teks adalah analisis sentimen, yaitu sebuah proses untuk menganalisa sekumpulan teks untuk mendapatkan informasi sentimen atau informasi emosi berdasarkan teks yang dituliskan oleh penulis [7]. Analisis sentimen banyak terdapat pada media sosial, di mana seseorang bebas menuliskan komentar atau artikel dalam bentuk teks tertulis untuk menanggapi sebuah berita, informasi, produk, layanan, atau hal-hal dalam bentuk lainnya yang dipublikasikan secara umum [8]. Tujuan dari analisis sentimen adalah untuk mengetahui respon atau perasaan emosional seseorang terhadap suatu hal yang ditanggapinya dalam bentuk teks. Secara umum, analisis sentimen memiliki tiga buah kategori sentimen dasar yaitu positif, negatif, dan netral. Salah satu contoh penerapan analisis sentimen adalah untuk mengetahui hasil *review* yang dituliskan oleh seseorang setelah membeli sebuah produk pada situs *e-commerce* dengan menggunakan *multiclassification model* [9]. Setiap orang tentunya memiliki pengalaman yang berbeda-beda terhadap sebuah produk yang dibeli. Perbedaan pengalaman tersebut tentunya akan menghasilkan respon sentimen yang berbeda-beda pula sehingga hal tersebut merupakan sesuatu yang menarik untuk diteliti. Berdasarkan penelitian tersebut, analisis sentimen dapat dilakukan dengan baik meskipun belum optimal karena tata bahasa yang digunakan pada teks-teks *review* memiliki struktur yang cukup kompleks. Contoh penelitian lain adalah identifikasi sentimen untuk mengetahui respon para penggemar olah raga sepak bola terhadap tim atau pertandingan sepak bola yang mereka tuliskan pada media sosial Twitter [10]. Dengan mengumpulkan seluruh data *twit* yang terkait dengan sepak bola, hasil pemrosesan dapat menggambarkan jenis sentimen yang terkandung pada setiap data *twit* yang telah dikumpulkan. Luaran yang dihasilkan dari penelitian tersebut dapat mengidentifikasi sentimen yang terdapat pada setiap *twit* secara efektif.

Kecerdasan buatan merupakan salah satu bidang ilmu komputer yang dapat membuat sebuah komputer menjadi lebih cerdas seperti manusia dalam melakukan sebuah pekerjaan [11].

Beberapa hal yang dapat dilakukan oleh sebuah komputer setelah ditanamkan algoritma kecerdasan buatan adalah mengenali sesuatu [12], membuat sebuah keputusan berdasarkan data-data [13], serta menganalisa sekumpulan data dan mengelompokkannya ke dalam beberapa bagian berdasarkan aturan tertentu [14]. *K-Nearest Neighbors* (*K-NN*) merupakan salah satu metode di bidang kecerdasan buatan yang banyak digunakan untuk keperluan klasifikasi data, baik data berupa citra [15, 16] atau teks [17]. Salah satunya adalah penelitian mengenai pendeteksian ketersediaan area parkir berdasarkan nilai *Moment Invariants* pada citra dengan menggunakan metode *K-NN* [18]. Penelitian ini bertujuan untuk membangun sebuah sistem yang dapat memberikan informasi slot parkir yang tersedia pada sebuah area dengan memanfaatkan kamera *CCTV* (*Closed-Circuit Television*) yang terpasang pada area tersebut. Selain untuk keperluan keamanan, kamera dapat *CCTV* dimanfaatkan untuk mendeteksi *marker* dengan simbol berbeda-beda yang terpasang di setiap lantai *slot* parkir. Apabila kamera *CCTV* dapat mendeteksi simbol pada *marker*, maka area slot parkir yang diasosiasikan terhadap simbol tersebut tidak sedang digunakan. Dengan demikian, sistem tersebut dapat dibangun dengan biaya yang cukup terjangkau, karena cukup memanfaatkan peralatan yang sudah ada yaitu kamera *CCTV*. Pada penelitian tersebut, metode *K-NN* digunakan untuk mendeteksi simbol-simbol unik yang terdapat pada *marker* dan dapat menghasilkan tingkat akurasi baik yaitu 91,94%. Selain itu, metode *K-NN* juga dapat digunakan untuk pemrosesan teks, salah satunya adalah pengklasifikasian dokumen-dokumen berbasis teks [19]. Sebelum diproses lebih lanjut, proses ekstraksi fitur dilakukan terhadap seluruh dokumen dengan menggunakan metode *Principal Component Analysis* (*PCA*) agar mendapatkan ciri inti dari setiap dokumen. Secara garis besar, proses klasifikasi dapat dilakukan dengan baik dan efektif.

*Dataset* merupakan sebuah komponen penting dalam bidang kecerdasan buatan, khususnya untuk digunakan pada tahap pelatihan maupun pengujian [20]. Pada penelitian terdahulu, pembentukan *dataset* pernah dilakukan untuk membangun sebuah basis data yang berisi topik-topik pada media sosial Twitter [21]. Pembentukan *dataset* dilakukan dengan cara mengkompilasi data-data *twit* dari akun-akun media massa berbasis daring. Dari *twit* yang terdapat pada akun-akun tersebut, dilakukan pengelompokan data secara otomatis dengan menggunakan metode *K-Means*. Metode tersebut bekerja dengan cara melihat kemiripan data *twit* berdasarkan perhitungan menggunakan *cosine similarity*. Penelitian lain terkait pembangunan *dataset* bertujuan untuk mempelajari pola serangan *Denial of Service* (*DoS*) terhadap jaringan NETwork [22]. Pembentukan *dataset* secara otomatis dianggap penting agar dapat menyesuaikan perubahan tren yang mungkin terjadi sehingga pihak pengelola tidak perlu terus-menerus memperbarui *dataset* secara manual apabila terdapat perubahan. *Dataset* yang dihasilkan pada penelitian tersebut mampu memberikan nilai akurasi sebesar 99.5% dalam mengenali serangan *DoS*.

**Tabel 1. Isi Dataset Awal**

| **Judul** | Pembentukan Dataset Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity | Deteksi Area Parkir Mobil Berbasis Marker Menggunakan Moment Invariants dan K-NN | An improved kNN text classification method | A simulation work for generating a novel dataset to detect distributed denial of service attacks on Vehicular Ad hoc NETwork systems |
|---|---|---|---|---|
| **Penulis** | K. A. Nugraha & D. Sebastian | K. A. Nugraha | F. Wang, Z. Liu & C. Wang | F. A. Alhaidari & A. M. Alrehan |
| **Tahun** | 2018 | 2019 | 2019 | 2021 |
| **Topik** | Pembentukan Dataset | Klasifikasi K-NN | Klasifikasi K-NN | Pembentukan Dataset |
| **Objek** | Teks Tweet (Twitter) | Citra | Dokumen Teks | Data Traffic Jaringan |

Berdasarkan pemaparan di atas, penelitian ini mencoba untuk melakukan pembentukan *dataset* untuk keperluan analisis sentimen, dalam kasus ini adalah untuk topik elektronik. Penelitian dilakukan dengan memanfaatkan metode *K-NN* ditambah dengan beberapa metode pra-pemrosesan seperti tokenisasi dan *stopwording*. Perbedaan mendasar antara penelitian ini dengan penelitian terdahulu dapat ditunjukkan pada Tabel 1. Luaran dari penelitian ini diharapkan dapat

mendukung penelitian-penelitian lain yang sejenis, khususnya untuk keperluan pengadaan *dataset* dalam bidang analisis sentimen.

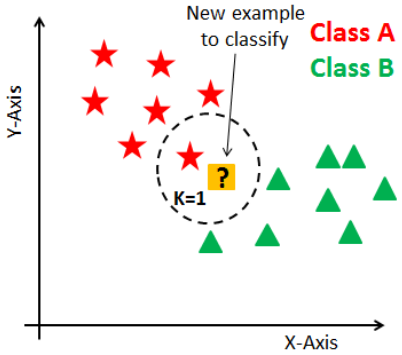
## 3. Metodologi Penelitian

Penelitian ini terdiri dari beberapa tahap yaitu pengumpulan data, perancangan struktur *dataset*, implementasi algoritma (*K-NN*), dan evaluasi. Tahap pengumpulan data dilakukan dengan memilih sebanyak lima akun *brand* elektronik Indonesia pada Instagram dengan jumlah *follower* terbanyak. Untuk setiap akun, diambil seluruh data komentar dari sepuluh *post* terakhir milik masing-masing akun. Dari proses tersebut, jumlah keseluruhan data yang berhasil dikumpulkan adalah sebanyak 10.351 komentar.

Setelah seluruh data komentar berhasil dikumpulkan, langkah berikutnya adalah melakukan perancangan struktur *dataset* guna menentukan basis data dan perangkat pemrosesan yang sesuai dengan karakteristik data yang telah berhasil dikumpulkan. Seluruh komentar akan disimpan menggunakan basis data berupa file teks standar dengan pertimbangan minimnya operasi *dataset* yang diperlukan serta tidak adanya parameter-parameter khusus yang harus diolah. Oleh sebab itu, penggunaan perangkat basis data yang lebih kompleks seperti *Structured Query Language* (*SQL*) hanya akan memperlambat pemrosesan tanpa berpengaruh pada hasil.

Algoritma diimplementasikan menggunakan bahasa pemrograman Python dengan pertimbangan ketersediaan pustaka yang cukup lengkap untuk keperluan pengolahan teks seperti tokenisasi dan *stopwording*. Mula-mula seluruh data komentar yang telah berhasil dikumpulkan sebelumnya akan dibaca terlebih dahulu. Kemudian, untuk setiap data komentar akan dilakukan proses tokenisasi untuk memotong sebuah kalimat menjadi kata-kata (*token*). Selain itu, akan dilakukan juga proses *stopwording*, yaitu proses untuk menghilangkan kata-kata yang kurang bermakna seperti kata tunjuk atau kata sambung. Pada permasalahan analisis sentimen, kata-kata tersebut akan dianggap sebagai kata yang tidak memiliki informasi sentimen apapun.

Pada tahap inisialisasi awal, *dataset* sentimen akan diisi menggunakan kata-kata positif dan negatif secara umum, contohnya kata bagus-jelek, baik-buruk, puas-kecewa, dan sejenisnya. Jumlah kata awal untuk *dataset* masing-masing jenis sentimen adalah sebanyak 15 kata. Langkah berikutnya adalah membandingkan setiap kata pada komentar tersebut dengan seluruh kata positif dan negatif yang terdapat pada *dataset*, kemudian menyimpulkan hasil sentimen yang terdapat pada kalimat tersebut dengan cara menghitung kata terbanyak antara kata positif atau negatif. Apabila jumlah kata positif dan negatif tidak ada sama sekali atau berjumlah sama, maka kalimat tersebut tidak akan diproses lebih lanjut. Setelah mendapatkan daftar komentar dengan label positif dan negatif tahap pertama, seluruh kata-kata unik pada komentar tersebut akan ditambahkan ke dalam *dataset* sentimen sesuai labelnya masing-masing. Langkah berikutnya adalah dengan menerapkan metode *K-NN* untuk mencari komentar-komentar lain yang belum berhasil diproses sebelumnya terhadap komentar-komentar yang telah memiliki label. *K-NN* bekerja dengan cara membandingkan data uji dengan sejumlah (K) data dari *dataset* yang memiliki tingkat kemiripan paling tinggi, kemudian menyimpulkan kelas berdasarkan kelas terbanyak yang muncul dari K-data tersebut. Ilustrasi ditunjukkan pada Gambar 1.




**Gambar 1. Ilustrasi cara kerja *K-NN*** [23]

Perhitungan kemiripan dilakukan dengan menggunakan persamaan *cosine similarity* seperti yang ditunjukkan pada Persamaan 1.

$$cos\_similarity(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

Keterangan:
n = jumlah atribut
x = data uji
y = data target

Sebagai contoh perhitungan di atas, terdapat kalimat uji 'fiturnya bagus dan keren parah' yang akan dibandingkan terhadap dua kalimat dari dataset yang berbeda, yaitu 'kameranya bagus dan keren' (positif) dan 'kualitas layarnya parah' (negatif). Pada contoh tersebut diasumsikan K bernilai satu sehingga hanya dibutuhkan satu kalimat yang memiliki nilai kemiripan paling tinggi untuk menentukan kelas dari kalimat uji. Pertama-tama, buat daftar kata hasil gabungan dari kalimat uji dengan kalimat positif. Kemudian, setiap kata yang ada di dalam sebuah kalimat akan diberi nilai 1. Hal tersebut ditunjukkan pada Tabel 2.

**Tabel 2. Isi Dataset Awal**

| | fiturnya | bagus | dan | keren | parah | kameranya |
|---|---|---|---|---|---|---|
| **kalimat uji** | 1 | 1 | 1 | 1 | 1 | 0 |
| **kalimat positif** | 0 | 1 | 1 | 1 | 0 | 1 |

Apabila dihitung menggunakan *cosine similarity*, maka tingkat kemiripan dari kedua kalimat tersebut adalah:

$$cos\_similarity(\mathbf{uji},\mathbf{positif}) = \frac{(1\,x\,0)+(1\,x\,1)+(1\,x\,1)+(1\,x\,1)+(1\,x\,0)+(0\,x\,1)}{\sqrt{1^2+1^2+1^2+1^2+1^2+0^2}\times\sqrt{0^2+1^2+1^2+1^2+0^2+1^2}} = \frac{3}{2.24\,x\,2} = \frac{3}{4.48} = 0.66$$

Lakukan langkah yang sama untuk kalimat negatif seperti ditunjukkan pada Tabel 3.

**Tabel 3. Isi Dataset Awal**

| | fiturnya | bagus | dan | keren | parah | kualitas | layarnya |
|---|---|---|---|---|---|---|---|
| **kalimat uji** | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| **kalimat positif** | 0 | 0 | 0 | 0 | 1 | 1 | 1 |

Apabila dihitung menggunakan *cosine similarity*, maka tingkat kemiripan dari kedua kalimat tersebut adalah:

$$cos\_similarity(\mathbf{uji},\mathbf{positif}) = \frac{(1\,x\,0)+(1\,x\,0)+(1\,x\,0)+(1\,x\,0)+(1\,x\,1)+(0\,x\,1)+(0\,x\,1)}{\sqrt{1^2+1^2+1^2+1^2+1^2+0^2+0^2}\times\sqrt{0^2+0^2+0^2+0^2+1^2+1^2+1^2}} = \frac{1}{2.24\,x\,1.73} = \frac{1}{3.87} = 0.25$$

Karena nilai kemiripan kalimat uji menggunakan *cosine similarity* terhadap kalimat positif lebih tinggi (0.66) dibandingkan dengan kalimat negatif (0.25), maka dapat disimpulkan bahwa kalimat uji 'fiturnya bagus dan keren parah' adalah kalimat positif.

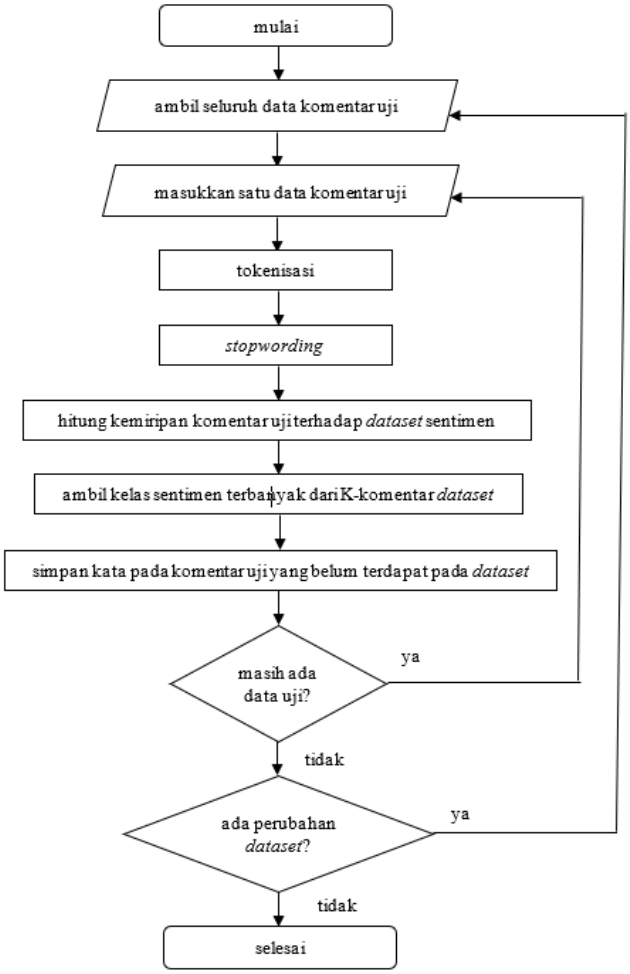
Berdasarkan hasil kemiripan sebuah komentar terhadap komentar-komentar yang berlabel, akan diambil K-komentar berdasarkan komentar dengan tingkat kemiripan paling tinggi. Komentar yang diujikan akan diberi label sesuai dengan kelas yang paling banyak muncul di antara K-komentar tersebut. Selanjutnya, seluruh kata yang terdapat pada komentar akan ditambahkan ke dalam *dataset* sesuai dengan label yang diberikan. Apabila sebuah kata telah terdapat dalam *dataset*, maka kata tersebut tidak akan ditambahkan sehingga dalam *dataset* hanya terdapat satu buah kata unik untuk setiap kata. Langkah berikutnya adalah melakukan penyaringan kata yang terdapat pada kedua label (positif dan negatif) *dataset*. Apabila sebuah kata terdapat dalam kedua basis data, maka kata tersebut akan dihapus karena dianggap sebagai kata yang tidak unik dan tidak mencerminkan sentimen apapun. Dengan demikian, setiap label pada *dataset* hanya mengandung kata-kata eksklusif yang tidak terdapat pada label lain. Kata-

kata di luar label *dataset* akan dianggap sebagai kata netral yang tidak memiliki unsur sentimen apapun. Apabila dalam sebuah iterasi terdapat perubahan jumlah data dalam *dataset*, maka iterasi akan diulang kembali untuk komentar-komentar yang belum berhasil dikenali sebelumnya. Iterasi akan berhenti saat jumlah data dalam *dataset* tidak mengalami perubahan lebih lanjut. Pengujian dilakukan dengan menghitung nilai akurasi untuk masing-masing percobaan nilai k, yaitu k = 1 s/d 10. Dari hasil *dataset* yang didapatkan, akurasi ketepatan *dataset* akan dihitung berdasarkan nilai *recall* menggunakan Persamaan 2.

$$recall = \frac{TP}{TP+FP} \tag{2}$$

Keterangan:
*TP* = True Positive
*FP* = False Positive

Untuk melihat alur kerja algoritma, diagram alir mengenai cara kerja algoritma secara keseluruhan dapat dilihat pada Gambar 2.




**Gambar 2. Diagram alir**

## 4. Hasil dan Diskusi

Pengujian dilakukan dengan menggunakan 10.351 data komentar dengan nilai parameter k = 1 s/d 10. Pada tahap awal, diperlukan *dataset* untuk inisialisasi saat pertama kali dijalankan. *Dataset* awal tersebut terdiri dari dua kelas, yaitu positif dan negatif, di mana masing-masing kelas berisi 10 kata sentimen secara umum sesuai kelas masing-masing. *Dataset* tersebut berisi

kata-kata positif dan negatif yang bersifat umum, tidak terkait dengan topik apapun termasuk topik elektronik yang menjadi obyek penelitian ini karena *dataset* tersebut bertujuan dapat digunakan untuk berbagai macam bidang atau topik. Daftar kata untuk *dataset* inisialisasi dapat dilihat pada Tabel 4.

**Tabel 4. Isi Dataset Awal**

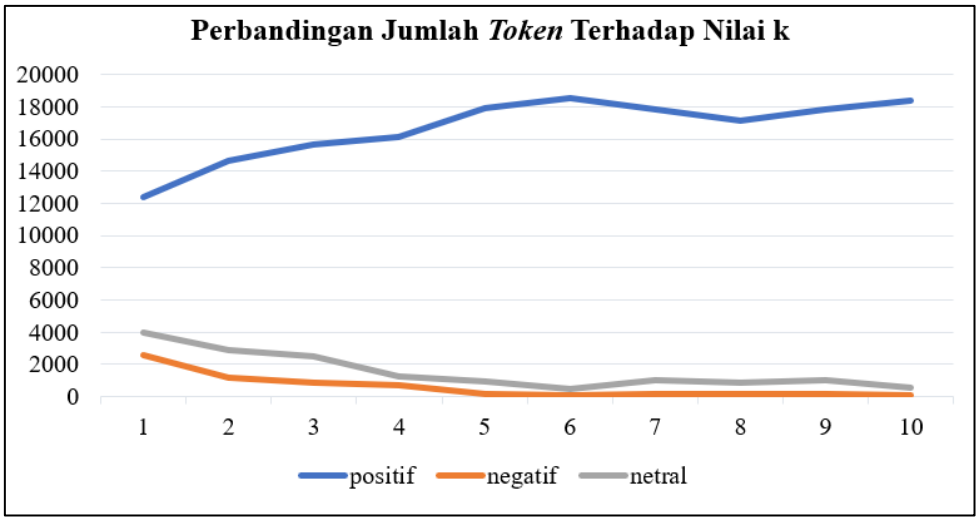
| Positif | Negatif |
|---|---|
| juara | jelek |
| bagus | payah |
| baik | buruk |
| rajin | malas |
| keren | kurang |
| mantap | susah |
| lancar | lambat |
| puas | kecewa |
| menang | kalah |
| jujur | bohong |

Berdasarkan hasil pengujian pembentukan *dataset* dengan metode *K-NN*, seluruh nilai *K* (1 s/d 10) menghasilkan nilai akurasi yang tidak jauh berbeda. Hal tersebut disebabkan karena metode *K-NN* dilakukan secara iteratif atau berulang hingga tidak ada lagi data komentar yang dapat diproses. Setelah melalui beberapa iterasi, seluruh nilai *K* dapat memproses seluruh data komentar yang ada. Dengan demikian, luaran akhir yang dihasilkan untuk masing-masing nilai *K* memiliki isi yang hampir serupa. Hasil pengujian nilai *K* secara keseluruhan dapat dilihat pada Tabel 5.

**Tabel 5. Jumlah Token *Dataset* Berdasarkan Nilai K**

| K | Epoch | Jumlah Komentar | Jumlah Token | | |
|---|---|---|---|---|---|
| | | | Positif | Negatif | Netral |
| 1 | 1 | 10351 | 12420 | 2610 | 3992 |
| 2 | 6 | 10019 | 14621 | 1205 | 2893 |
| 3 | 1 | 10351 | 15669 | 836 | 2517 |
| 4 | 7 | 9978 | 16133 | 745 | 1273 |
| 5 | 1 | 10351 | 17917 | 152 | 953 |
| 6 | 4 | 10349 | 18522 | 52 | 448 |
| 7 | 1 | 10351 | 17829 | 175 | 1018 |
| 8 | 7 | 9613 | 17142 | 166 | 867 |
| 9 | 1 | 10351 | 17815 | 182 | 1025 |
| 10 | 6 | 10245 | 18407 | 63 | 548 |

Secara umum, semakin besar nilai *K*, semakin banyak *token* berlabel positif yang dihasilkan. Sebaliknya, jumlah *token* negatif dan netral akan semakin sedikit seiring bertambahnya nilai *K*. Berdasarkan hal tersebut, dapat disimpulkan bahwa komentar-komentar yang terdapat pada akun Instagram yang digunakan pada penelitian ini didominasi oleh komentar-komentar yang memiliki sentimen positif. Pada *K* dengan nilai ganjil, *epoch* dapat dipastikan bernilai satu. Hal tersebut dikarenakan hanya terdapat dua jenis kelas yaitu positif dan negatif. Oleh karena itu, apabila *K* bernilai ganjil maka hasil untuk kedua kelas tidak mungkin sama. Dengan kata lain, dapat dipastikan bahwa salah satu kelas akan lebih dominan dibandingkan dengan kelas yang lain. Grafik perbandingan jumlah *token* terhadap nilai *K* terdapat pada Gambar 3.




**Gambar 3. Grafik Perbandingan Jumlah *Token* Terhadap Nilai *k***

Pada Tabel 6, dapat dilihat nilai akurasi berdasarkan $K$ = 1 s/d 10 untuk setiap label. Sebelum menghitung nilai akurasi, seluruh kata pada label yang berupa kata benda, nama (orang, kota, dan sejenisnya), angka, serta kata-kata yang mengandung kesalahan penulisan (*typo*) dihilangkan terlebih dahulu dari *dataset* karena dianggap sebagai data yang tidak valid. Secara keseluruhan, nilai akurasi terbaik dihasilkan dengan parameter $K$ = 1.

**Tabel 6. Hasil Akurasi Berdasarkan Nilai K**

| K | Akurasi (*Recall*) | | |
|---|---|---|---|
| | **Positif** | **Negatif** | **Netral** |
| 1 | 33.38% | 59.96% | 56.60% |
| 2 | 23.33% | 43.32% | 40.01% |
| 3 | 26.67% | 30.02% | 33.35% |
| 4 | 33.34% | 44.56% | 44.30% |
| 5 | 26.67% | 46.71% | 46.64% |
| 6 | 30.01% | 42.31% | 33.33% |
| 7 | 33.02% | 53.14% | 43.36% |
| 8 | 30.50% | 43.98% | 39.91% |
| 9 | 36.67% | 35.71% | 46.63% |
| 10 | 34.49% | 41.27% | 46.90% |

Pada *dataset* berlabel positif, penambahan nilai $K$ berpengaruh terhadap bertambahnya jumlah *token* yang dihasilkan. Sementara itu, penambahan jumlah *token* pada *dataset* positif tidak diimbangi dengan adanya *token-token* yang menjadi ciri dari sentimen positif. Dengan kata lain, *token-token* yang bertambah tersebut tidak dapat menjadi *dataset* positif. Hal tersebut berakibat pada menurunnya nilai akurasi *dataset*. Sementara itu, pada *dataset* negatif dan netral memiliki kecenderungan berkurangnya jumlah *token* seiring bertambahnya nilai $K$. Hal tersebut mengakibatkan terlalu sedikit *token* yang terdapat pada *dataset* tersebut sehingga nilai akurasi juga turut berkurang saat nilai $K$ semakin besar. Beberapa contoh kata untuk setiap label setelah dilakukan proses validasi dapat dilihat pada Tabel 7.

**Tabel 7. Contoh Isi *Dataset***

| No | Positif | Negatif | Netral |
|---|---|---|---|
| 1 | jujur | payah | baca |
| 2 | okay | geram | hadir |
| 3 | adil | gagu | datang |

| | | | |
|---|---|---|---|
| 4 | terkenal | ngaco | nonton |
| 5 | lebar | ngestuck | tingkat |
| 6 | komitmen | closed | ganti |
| 7 | powerful | jauh | sebentar |
| 8 | spektakuler | meledak | terlebih |
| 9 | adem | meresahkan | fokus |
| 10 | tenteram | gila | mendengar |

Beberapa kendala yang dihadapi oleh program adalah terdapat banyak kata pada komentar yang tidak dapat dikenali karena kesalahan penulisan seperti 'carjer' yang seharusnya adalah 'charger', penggunaan huruf yang tidak standar seperti 'temen' yang seharusnya adalah 'teman', serta penggunaan huruf yang berulang seperti 'kereeennn' yang seharusnya adalah 'keren'. Selain itu, terdapat banyak komentar yang mengandung ekspresi sentimen dengan menggunakan *emoticon* / *emoji*, di mana komentar jenis ini tidak dapat diproses oleh program karena tidak dapat mengenali kode atau format *emoticon* / *emoji*.

## 5. Kesimpulan dan Saran

Secara keseluruhan, *dataset* yang berhasil dibentuk memiliki tingkat akurasi tertinggi sebesar 33.38% (positif), 59.96% (negatif), dan 56.60% (netral) untuk $K = 1$. Beberapa kendala yang dihadapi terkait dengan pemrosesan kata disebabkan karena cukup banyak terdapat kata-kata dalam bentuk tidak baku atau tidak standar, seperti singkatan, huruf yang berulang, serta terdapat adanya kesalahan penulisan kata (*typo*). Pada penelitian selanjutnya, pemrosesan teks dengan sumber data yang didapat dari media sosial, khususnya dalam bentuk komentar bebas, sebaiknya ditambahkan proses perbaikan kata atau normalisasi terlebih dahulu. Hal ini bertujuan agar kata-kata yang mengandung kesalahan dapat diperbaiki terlebih dahulu guna mempermudah pemrosesan teks pada tahap berikutnya. Selain itu, algoritma hendaknya juga memiliki kemampuan untuk memproses *emoticon* / *emoji*, dikarenakan konten yang ada pada media sosial saat ini seringkali menggunakan *emoticon* / *emoji* untuk mengekspresikan pesan yang hendak disampaikan.

## Referensi

[1] M. Iqbal, "Instagram Revenue and Usage Statistics (2021)," Business of Apps, 8 Maret 2021. [Online]. Available: https://www.businessofapps.com/data/instagram-statistics/. [Diakses 14 Maret 2021].

[2] H. Tankovska, "Distribution of Instagram users worldwide as of January 2021, by age group," Statista, 10 Februari 2021. [Online]. Available: https://www.statista.com/statistics/325587/instagram-global-age-group/. [Diakses 12 Maret 2021].

[3] Y. Yao, J. Zhang, F. Shen, L. Liu, F. Zhu, D. Zhang dan H. T. Shen, "Towards Automatic Construction of Diverse, High-Quality Image Datasets," *IEEE Transactions on Knowledge and Data Engineering,* vol. 32, no. 6, pp. 1199 - 1211, Mar. 2019.

[4] V. Vonikakis, R. Subramanian dan S. Winkler, "Shaping datasets: Optimal data selection for specific target distributions across dimensions," dalam *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, Sept. 2016.

[5] X. Tang, Z. Huang, D. Eyers, S. Mills dan M. Guo, "Scalable Multicore k-NN Search via Subspace Clustering for Filtering," *IEEE Transactions on Parallel and Distributed Systems,* vol. 26, no. 12, pp. 3449 - 3460, Nov. 2014.

[6] K. A. Nugraha dan D. Sebastian, "Analisis Trend Akun Media Sosial Twitter Menggunakan TF-IDF dan Cosine Similarity," dalam *Seminar Nasional ReTII ke-13*, Yogyakarta, Indonesia, Nov. 2018.

[7] H. T. Phan, V. C. Tran, N. T. Nguyen dan D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," *IEEE Access,* vol. 8, pp. 14630 - 14641, Jan. 2020.

[8] L. Yang, Y. Li, J. Wang dan R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," *IEEE Access,* vol. 8, no. IEEE, pp. 23522 - 23530, Jan. 2020.

[9] S. Zhang, D. Zhang, H. Zhong dan G. Wang, "A Multiclassification Model of Sentiment for E-Commerce Reviews," *IEEE Access,* vol. 8, pp. 189513 - 189526, Oct. 2020.

[10] S. Aloufi dan A. E. Saddik, "Sentiment Identification in Football-Specific Tweets," *IEEE Access,* vol. 6, pp. 78609 - 78621, Dec. 2018.

[11] P.-H. Lin, A. Wooders, J. T.-Y. Wang dan W. M. Yuan, "Artificial Intelligence, the Missing Piece of Online Education?," *IEEE Engineering Management Review,* vol. 46, no. 3, pp. 25 - 28, Oct. 2018.

[12] J. Kong, M. Chen, M. Jiang, J. Sun dan J. Hou, "Face Recognition Based on CSGF(2D)2PCANet," *IEEE Access,* vol. 6, pp. 45153 - 45165, Sept. 2018.

[13] M. C. Canellas, K. M. Feigh dan Z. K. Chua, "Accuracy and Effort of Decision-Making Strategies With Incomplete Information: Implications for Decision Support System Design," *IEEE Transactions on Human-Machine Systems ,* vol. 45, no. 6, pp. 686 - 701, Jul. 2015.

[14] X. Wang dan H. Ji, "Semi-Supervised Hyperspectral Image Classification Based on Label Propagation via Selected Path," *IEEE Access ,* vol. 8, pp. 221225 - 221234, Dec. 2020.

[15] K. A. Nugraha, W. Hapsari dan N. A. Haryono, "Analisis Tekstur Pada Citra Motif Batik Untuk Klasifikasi Menggunakan K-NN," *Informatika: Jurnal Teknologi Komputer dan Informatika,* vol. 10, no. 2, pp. 135-140, Apr. 2014.

[16] L. A. Sunjoyo, R. G. Santosa dan K. A. Nugraha, "Implementasi Transformasi Haar Wavelet untuk Deteksi Citra Jeruk Nipis yang Busuk," *Informatika: Jurnal Teknologi Komputer dan Informatika,* vol. 12, no. 2, pp. 165-173, Nov. 2016.

[17] K. A. Nugraha dan Herlina, "Klasifikasi Pertanyaan Bidang Akademik Berdasarkan 5W1H menggunakan K-Nearest Neighbors," *JEPIN (Jurnal Edukasi dan Penelitian Informatika),* vol. 7, no. 1, pp. 44-51, Apr. 2021.

[18] K. A. Nugraha, "Deteksi Area Parkir Mobil Berbasis Marker Menggunakan Moment Invariants dan K-NN," *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI),* vol. 5, no. 1, pp. 112-121, May. 2019.

[19] F. Wang, Z. Liu dan C. Wang, "An improved kNN text classification method," *International Journal of Computational Science and Engineering (IJCSE),* vol. 20, no. 3, Nov. 2019.

[20] F. Iglesias, T. Zseby, D. Ferreira dan A. Zimek, "MDCGen: Multidimensional Dataset Generator for Clustering," *Journal of Classification,* vol. 36, p. 599–618 , Apr. 2019.

[21] K. A. Nugraha dan D. Sebastian, "Pembentukan Dataset Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & Cosine Similarity," *Jurnal Teknik Informatika dan Sistem Informasi (JuTISI),* vol. 4, no. 3, pp. 376-386, Dec. 2018.

[22] F. A. Alhaidari dan A. M. Alrehan, "A simulation work for generating a novel dataset to detect distributed denial of service attacks on Vehicular Ad hoc NETwork systems," *International Journal of Distributed Sensor Networks,* vol. 17, no. 3, Mar. 2021.

[23] A. Navlani, "KNN Classification using Scikit-learn," Datacamp, 2 Agustus 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. [Diakses 24 April 2021].