PROSIDING Seminar Nasional SISFOTEK ISSN 2597-3584 (Media Online)
Sistem Informasi dan Teknologi
website : www.seminar.iaii.or.id | email : seminar@iaii.or.id

Data Mining Untuk Klasifikasi Produk Menggunakan Algoritma K-Nearest Neighbor Pada Toko Online

Ma'ruf Aziz Muzani[1], M. Iqbal Abdullah Sukri[2], Syifa Nur Fauziah[3], Agus Fatkhurohman[4] , Dhani Ariatmanto[5]
[1,2,3,4,5] Master Teknik Informatika, Universitas Amikom Yogyakarta
ma'ruf_muzani@students.amikom.ac.id

Abstract

The rapid growth of e-commerce in Indonesia has been largely facilitated by the presence of e-marketplaces. The e-marketplace trend in Indonesia continues to develop along with the development of technology and the internet. During its development, e-marketplaces offer more and more products. As a result, buyers need more effort to find the product they want. In order to facilitate the search for these products, a product classification is carried out. This study classifies products in the Shopee emarketplace using the K-Nearest Neighbor algorithm. The product data used comes from web scraping in the categories of cellphones and accessories, Muslim fashion, and home appliances. The stages of the classification system begin with the preprocessing stage, then the term weighting stage uses the TF-IDF method, then cosine similarity to calculate the similarity distance between documents, and then sorting the results of the cosine similarity to retrieve data for the number of k values. Based on testing on 9 product data with three different k values. Obtained an average that shows the lowest accuracy, precision, and recall results when the value of k = 3. The accuracy result is 88.89%, precision is 83.33%, and a recall of 100% is obtained when using the value of k = 5 or k = 7.

Keywords - Classification, E-marketplace, K-Nearest Neighbor, Text Mining

Abstrak

Pesatnya pertumbuhan e-commerce di Indonesia sebagian besar telah difasilitasi oleh kehadiran e-marketplace. Tren e-marketplace di Indonesia terus berkembang seiring dengan perkembangan teknologi dan internet. Dalam perkembangannya, e-marketplace menawarkan produk yang semakin banyak. Akibatnya, pembeli membutuhkan lebih banyak upaya untuk menemukan produk yang mereka inginkan. Untuk memudahkan pencarian produk tersebut maka dilakukan klasifikasi produk. Penelitian ini mengklasifikasikan produk di emarketplace Shopee menggunakan algoritma K-Nearest Neighbor. Data produk yang digunakan berasal dari web scraping kategori handphone dan aksesoris, busana muslim, dan peralatan rumah tangga. Tahapan sistem klasifikasi dimulai dengan tahap preprocessing, kemudian tahap term weighting menggunakan metode TF-IDF, kemudian cosinus similarity untuk menghitung jarak kemiripan antar dokumen, kemudian mengurutkan hasil cosinus similarity untuk mengambil data bilangan. dari nilai k. Berdasarkan pengujian pada 9 data produk dengan tiga nilai k yang berbeda. Diperoleh rata-rata yang menunjukkan hasil akurasi, presisi, dan recall terendah saat nilai k = 3. Hasil akurasi 88,89%, presisi 83,33%, dan recall 100% diperoleh saat menggunakan nilai k = 5 atau k = 7.

Kata Kunci - Klasifikasi, E-marketplace, K-Nearest Neighbor, Text Mining

1. Pendahuluan

Banyaknya pengguna internet dan media sosial di Indonesia menyimpan potensi besar yang bisa dimanfaatkan untuk melakukan transaksi penjualan secara *online*. Hal tersebut sejalan dengan meningkatnya minat berbelanja masyarakat Indonesia, yang didukung oleh perubahan perilaku konsumen untuk melakukan transaksi digital. GlobalWebIndex [1] melaporkan bahwa Indonesia memiliki tingkat tertinggi penggunaan *e-commerce* di antara negara manapun di dunia, dengan 90 persen pengguna internet di negara tersebut yang berusia antara 16 dan 64 tahun melaporkan bahwa mereka telah membeli produk dan layanan secara online. Aktivitas *e-commerce* di Indonesia dapat dilihat pada Gambar 1.

Gambar 1. *E-commerce Activities In Indonesia*

Hadirnya *e-commerce* memungkinkan masyarakat untuk berbelanja dimana saja dan kapan saja melalui media internet. Salah satu model bisnis *e-commerce* yaitu *e-marketplace*. *E-marketplace* merupakan sebuah website atau aplikasi jual beli online yang memiliki banyak *vendor*, model bisnis yang digunakan adalah *Customer to Customer* (C2C). Gambar 2 merupakan data transaksi *e-commerce* di Indonesia berdasarkan kategori. Menurut data awal tahun 2019 tersebut, pembelanjaan *online* konsumen di Indonesia paling banyak pada kategori *travel*, perangkat elektronik, dan *fashion* [1].

Gambar 2. *E-commerce Spend By Category*

Dengan semakin bertambahnya jumlah penjual dan pembeli pada *e-marketplace* menciptakan persaingan yang ketat antar penjual, sehingga penjual harus mencari cara agar produknya menarik dan dibeli. Salah satunya, dengan cara membuat judul produk yang unik, seperti menambahkan kata gratis, garansi, dan lain sebagainya. Selain itu, penjual melakukan penambahan merek terkenal lain di belakang judul produk untuk meningkatkan peringkat pada proses pencarian. Di sisi lain, calon pembeli mengalami kesulitan untuk mencari produk yang sesuai dengan keinginannya karena banyaknya produk yang dijual [2].

Text mining merupakan proses ekstraksi informasi yang berguna dari dokumen-dokumen teks tak terstruktur (*unstructured*) [3]. *Text mining* dapat digunakan untuk mengklasifikasikan dokumen ke dalam suatu kategori. Tujuan dari kategorisasi teks adalah menguji pengklasfikasian dokumen yang belum diketahui kategorinya, sehingga apabila terdapat dokumen baru akan lebih mudah diklasifikasikan pada suatu kategori berdasarkan dokumen-dokumen yang ada [4]. Salah satu pemanfaatan *text mining* pada *e-marketplace* yaitu untuk mengelompokkan data produk yang sama agar pembeli lebih mudah ketika melakukan pencarian produk yang diinginkan.

Penelitian yang dilakukan oleh Danny Sebastian pada tahun 2019 dengan judul "Implementasi Algoritme *KNearest Neighbor* untuk Melakukan Klasifikasi Produk dari beberapa *E-marketplace*" [2]. Pada penelitian tersebut metode *K-Nearest Neighbor* dapat melakukan klasifikasi produk dari *e-marketplace*, khususnya tokopedia dan bukalapak. Berdasarkan pengujian 1, disimpulkan bahwa nilai k yang optimal adalah 5. Kemudian akurasi yang dihasilkan pada pengujian 2 adalah 96,67%. Nilai akurasi ini dapat dikatakan baik karena melebihi 90%.

Penelitian oleh Ivan Jaya, Ainul Hizriadi, Evi Sersanti Purba pada tahun 2018 yang berjudul "Klasifikasi Surat Laporan Kehilangan Kepolisian Menggunakan Algoritme *K-Nearest Neighbor*" [5]. Memberikan kesimpulan algoritme *K-Nearest Neighbor* dapat diterapkan untuk mengklasifikasi surat laporan kehilangan kepolisian dengan nilai akurasi sebesar 91,75% dan mampu mengelompokkan isi surat yang memiliki kesamaan *term*.

Penelitian yang dilakukan oleh Difari Afreyna Fauziah, dkk pada tahun 2018 yang berjudul "Klasifikasi Berita Politik Menggunakan Algoritme *K-Nearest Neighbor"* [6], bertujuan untuk mencari nilai akurasi algoritme *K-Nearest Neighbor* terhadap objek yeng mempunyai nilai *similarity* tinggi. Objek yang digunakan pada penelitian tersebut menggunakan berita yang berkategori politik. Hasil pengujian menunjukkan bahwa algoritme *KNearest Neighbor* bekerja dengan baik sehingga algoritme *K-Nearest Neighbor* cocok dapat direkomendasikan untuk proses klasifikasi suatu dokumen dengan nilai *similarity* yang tinggiPenelitian oleh Danny Sebastian pada tahun 2017 yang berjudul "Rancang Bangun *Website* Klasifikasi Untuk Pencarian Produk Pasar Online Menggunakan Algoritme *K-Nearest Neighbor*" [7]. Penelitian ini menghitung nilai kecenderungan untuk data uji ke masing-masing kelas, nilai k yang digunakan yaitu k = 5. Berdasarkan nilai kecenderungan, nilai tertinggi adalah 1.2421 yang dimiliki oleh kelas Oppo Find 5 Mini. Dokumen uji diklasifikasikan ke kelas Oppo Find 5 Mini, sehingga hasil klasifikasi menggunakan perhitungan tersebut sesuai dengan seharusnya dimana dokumen uji masuk ke kelas Oppo Find 5 Mini.

Penelitian dengan judul "Implementasi Algoritme *KNearest Neighbor Dalam Pengklasifikasian Follower Twitter yang Menggunakan Bahasa Indonesia*" yang dilakukan oleh Muhammad Rivki, dan Adam Mukharil Bachtiar pada tahun 2017 [8], menghasilkan nilai akurasi terbesar untuk empat kali pengujian sebesar 68 %. Aplikasi dapat membantu membantu pengguna twitter untuk melakukan promosi terhadap *follower* yang sudah diklasifikasikan.

Penelitian yang terkait dengan latar belakang permasalahan diatas sudah pernah dilakukan oleh Danny Sebastian (2019) dengan judul Implementasi Algoritma *K-Nearest Neighbor* untuk Melakukan Klasifikasi Produk dari beberapa *E-marketplace* [2]. Pada penelitian tersebut metode *K-Nearest Neighbor* dapat melakukan klasifikasi produk dari *e-marketplace*, khususnya tokopedia dan bukalapak. Berdasarkan pengujian 1, disimpulkan bahwa nilai k yang optimal adalah 5. Kemudian akurasi yang dihasilkan pada pengujian 2 adalah 96,67%. Nilai akurasi ini dapat dikatakan baik karena melebihi 90%.

Berdasarkan penelitian sebelumnya, maka dalam penelitian ini penulis melanjutkan saran pengembangan dengan melakukan pengujian ke data

produk yang berasal dari *e-marketplace* selain tokopedia dan bukalapak. Klasifikasi yang dilakukan pada penelitian ini menggunakan metode *K-Nearest Neighbor*, sehingga klasifikasi dilakukan berdasarkan data *training* dilihat dari jarak yang paling dekat dengan objek berdasarkan nilai k.

Berdasarkan latar belakang yang telah dikemukakan, maka permasalahan yang dapat dirumuskan adalah "Bagaimana cara mengimplementasikan algoritme *KNearest Neighbor* untuk melakukan klasifikasi produk pada *e-marketplace* Shopee?".

Berdasarkan permasalahan di atas, tujuan yang ingin dicapai dalam penelitian ini yaitu, mengklasifikasikan produk pada *e-marketplace* Shopee menggunakan algoritme *K-Nearest Neighbor* dan menemukan tingkat akurasi algoritme *K-Nearest Neighbor* dalam mengklasifikasikan produk pada *emarketplace* Shopee.

2. Metode Penelitian

Dalam memperoleh data dan informasi sebagai penunjang penelitian ini, dilakukan pengambilan data dari suatu *website* secara spesifik. Dalam penelitian ini, *dataset* yang digunakan merupakan hasil *web-scraping* dari *website* Shopee.

Metode pengumpulan data dan informasi yang diperoleh dari buku-buku, jurnal ilmiah, tesis, internet, literatur, serta sumber-sumber lain yang berhubungan dengan objek penelitian. Peneliti melakukan studi pustaka mengenai metode dan algoritme yang akan digunakan, yaitu metode TF-IDF, *cosine similarity*, *confusion matrix*, dan algoritme *K-Nearest Neigbor*.

Metode analisis merupakan tahapan analisis terhadap studi litertaur yang bertujuan meningkatkan dan memperdalam pemahaman mengenai metode yang digunakan, yaitu algoritme *K-Nearest Neigbor* dan *text mining* untuk menyelesaikan permasalahan dalam klasifikasi produk.

Metode perancangan yang dilakukan yaitu bagaimana alur sistem ini akan berjalan. Selain itu, akan dilakukan perancangan *input* dan *output* sistem.

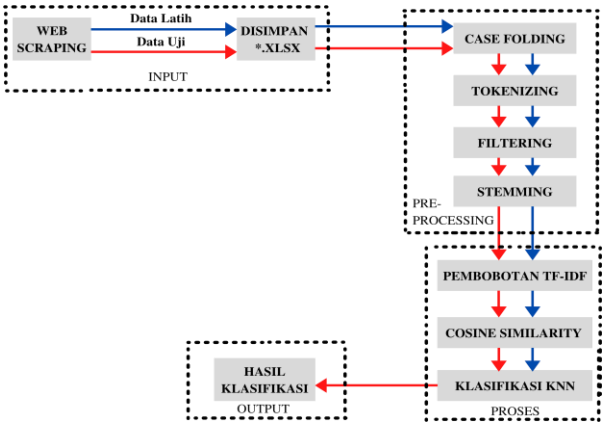
Dalam metode implementasi ini terdapat beberapa tahapan yang akan dilakukan yaitu, data hasil *scraping* akan digunakan sebagai data *training* dan data *testing*, *text preprocessing* dimana tahapan dari sistem klasifikasi dimulai dengan *preprocessing*, data produk yang telah dikumpulkan akan melewati empat tahap yaitu *case folding*, *tokenizing*, *stemming*, dan *filtering*. Selanjutnyap *t*embobotan TF-IDF, term-term yang terdapat pada setiap dokumen hasil *preprocessing* diberi nilai atau bobot menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*), dimana nilai tersebut akan dijadikan *input* pada klasifikasi. Setelah itu, klasifikasi algoritme *K-Nearest Neighbor* Setelah diperoleh bobot setiap term, kemudian dihitung jarak kemiripan antar dokumen menggunakan algoritme *cosine similarity*. Selanjutnya, data *training* diurutkan berdasarkan hasil *cosine similarity* untuk diambil sejumlah nilak k data yang memiliki kedekatan. Beriutnya pengujian sistem untuk menentukan tingkat akurasi algoritme *K-Nearest Neighbor* pada sistem ini akan dihitung menggunakan rumus *confusion matrix.*

Pada tahap selanjutnya yaitu metode pengujian dilakukan terhadap sistem untuk mengetahui seberapa baik sistem dalam melakukan klasifikasi, serta mengukur tingkat akurasi, nilai *precision*, dan nilai *recall* dari algoritme *K-Nearest Neighbor* didasarkan pada perhitungan *confusion matrix*.

3. Hasil dan Pembahasan

Secara garis besar alur penelitian pada penelitian ini yaitu pengumpulan *dataset* dari *website* Shopee. Setelah data diperoleh, kemudian diolah dengan tahapan *preprocessing* agar menghasilkan term-term yang bisa mewakili dokumen. Kemudian term-term tersebut dilakukan pembobotan dengan menggunakan metode TF-IDF. Hasil pembobotan TF-IDF digunakan untuk menghitung jarak kemiripan antara dokumen uji dengan dokumen latih menggunakan rumus *cosine similarity*. Setelah diperoleh hasil *cosine similarity* kemudian dilakukan klasifikasi dokumen menggunakan algoritme *K-Nearest Neighbor*. Algoritme *K-Nearest Neighbor* merupakan algoritme klasifikasi objek berdasarkan data pembelajaran yang jarak kemiripannya paling dekat dengan objek tersebut, sehingga diperlukan algoritme *cosine similarity* untuk melakukan perhitungan jarak.

Berikut adalah gambar arsitektur umum yang menggambarkan setiap tahapan yang digunakan pada penelitian ini, ditunjukkan pada Gambar 3.



Gambar 3. Arsitektur Umum

Tahapan – tahapan pada Gambar 3 menjelaskan tahapan yang pertama yaitu input pada sistem ini merupakan hasil dari *web scraping e-marketplace* Shopee dalam format .xlsx. Dokumen input dibagi menjadi data uji dan data latih. Data latih adalah dokumen yang sudah diberi label secara manual,

sedangkan data uji adalah dokumen yang belum memiliki label.

Selanjutnya p*re-processing*, ini terbagi dalam beberapa tahap diantaranya case folding yaitu proses untuk mengubah semua huruf kapital dalam teks menjadi huruf kecil (*lowercase*). Karakter lain yang bukan termasuk huruf dan angka dihilangkan. Selanjutnya *tokenizing* adalah proses pemotongan teks menjadi sebuah *token*, dapat berupa huruf, kata, dan angka. Kemudian tahapan f*iltering* adalah proses menghapus *stopword* atau kata yang kurang penting. Pada penelitian ini, penulis menggunakan *stopword* yang berasal dari Tala, F.Z dan menambahkan beberapa *stopword* yang berhubungan dengan data yang dikumpulkan. Seperti kata "original", "free", "garansi", "resmi", "terbaru", dan lain-lain. Beriutnya tahapan s***temming*** yaitu proses menghilangkan imbuhan pada token sehingga diperoleh kata dasar dengan menggunakan aturan-aturan tertentu.

Tahapan berikutnya dari gabar 3 yaitu Proses. Setelah tahap *pre-processing*, tahapan-tahapan selanjutnya diantaranya pembobotan TF-IDF yaitu proses pemberian nilai terhadap setiap term pada dokumen yang telah dilakukan *pre-processing*. Pada penelitian ini pembobotan terhadap term menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*). Metode TF-IDF merupakan metode yang umum digunakan untuk pemberian bobot suatu term. Kemudian ada tahapan c*osine similarity* merupakan proses untuk menghitung kemiripan antar dokumen, yaitu kemiripan vektor antara dokumen uji dengan setiap dokumen pada data latih. Selanjutnya *K-Nearest Neighbor* adalah algoritme yang digunakan untuk klasifikasi data uji dengan data yang paling mirip pada data latih yang ada. Nilai k pada algoritme *K-Nearest Neighbor* adalah k data terdekat dari data uji.

Tahapan akhir yaitu *output* yaitu hasil keluaran dari sistem ini berupa klasifikasi produk dari dokumen yang dimasukan. Proses klasifikasi telah selesai apabila telah ditentukan sebuah kelas untuk dokumen tersebut.

Data yang digunakan dalam penelitian ini adalah data kategori *handphone* dan aksesoris, fashion muslim, dan perlengkapan rumah yang diperoleh dari *website emarketplace* Shopee dengan cara *web scraping*.

Kategori tersebut dipilih dari 10 kategori produk terlaris di Shopee berdasarkan Team Asosiasi Digital Marketing Indonesia (www.digimind.id) yang telah melakukan riset pada 2,3 juta produk yang di jual di Shopee sampai dengan tanggal 7 Januari 2020 [9].

Peneliti menggunakan 15 data produk dari masingmasing kategori, sehingga dikumpulkan 45 data produk. Data produk dari masing-masing kategori kemudian dibagi menjadi 80% data latih dan 20% data uji.

Setelah data berhasil dikumpulkan, tahap selanjutnya adalah membuat kelas. Pada kategori handphone dan aksesoris, penulis memilih merek Iphone. Kemudian berdasarkan merek yang dipilih, dibuat kelas klasifikasi yang berasal dari nama produk. Adapun nama-nama kelas klasifikasi yang dibuat yaitu Iphone 7 Plus, dan Iphone X.

Pada kategori *fashion* muslim, penulis membuat nama-nama kelas klasifikasi yaitu Gamis, dan Dress. Selanjutnya, pada kategori perlengkapan rumah penulis membuat nama-nama kelas klasifikasi yaitu Rak Gantung, dan Hanger.

Pada tahap ini, penulis juga melakukan pelabelan secara manual terhadap 45 data produk yang sudah dikumpulkan. Pelabelan ini digunakan untuk membandingkan apakah hasil klasifikasi sistem sesuai dengan yang seharusnya.

Hasil klasifikasi sistem kemudian dievaluasi menggunakan rumus *confusion matrix* dengan menghitung nilai *precision*, *recall*, dan *accuracy*. Tahapan pengujian dilakukan terhadap 9 data uji. Data uji tersebut terdiri dari 3 data produk dari masing-masing kategori. Pengujian masing-masing kategori dilakukan dengan nilai k=3, k=5, dan k=7. Pemilihan nilai k berupa angka ganjil yaitu agar tidak terjadi hasil sama atau *draw*.

4. Kesimpulan

Berdasarkan penelitian tentang Penerapan Algoritme KNN dalam melakukan Klasifikasi Produk pada toko online yang telah dilakukan, maka dapat diperoleh kesimpulan bahwa dari hasil pengujian terhadap tiga kategori produk, diperoleh rata-rata nilai *accuracy*, *precision*, dan *recall* terendah ketika menggunakan nilai k=3. Kemudian nilai *accuracy*, *precision*, dan *recall* tertinggi diperoleh ketika menggunakan nilai k=5 atau k=7. Algoritme *K-Nearest Neighbor* dapat diimplementasikan pada penelitian ini dan memiliki hasil yang bagus karena tidak hanya nilai *accuracy* nya yang tinggi, tetapi juga nilai *precision* dan *recall.* Semakin banyak *dataset* yang digunakan, waktu yang dibutuhkan untuk klasifikasi semakin lama karena diperlukan perhitungan jarak dari setiap data uji pada keseluruhan data latih.

Daftar Rujukan

[1] S. Kemp and S. Moey, "Datareportal," 18 September 2019. [Online]. Available: https://datareportal.com/reports/digital-2019ecommerce-in-indonesia. [Accessed 17 May 2020].

[2] D. Sebastian, "Implementasi Algoritma K-Nearest Neighbor Untuk Melakukan Klasifikasi Produk dari Beberapa E-marketplace," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 5 Nomor 1 April 2019, pp. 51-61, 2019.

[3] A.-H. Tan, "Text Mining: The state of the art and challenges," *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, vol. 8, pp. 65-70, 1999.

[4] S. Gaikwad, A. Chaugule and P. Patil, "Text Mining Methods and Techniques," *International Journal of Computer Applications*, vol. 85, pp. 42-45, 2014.

[5] I. Jaya, A. Hizriadi and E. Purba, "Klasifikasi Surat Laporan Kehilangan Kepolisian Menggunakan Algoritma K-Nearest Neighbor," *TECHSI*, vol. 10, pp. 121-128, 2018.

[6] D. Fauziah, M. A and N. I, "Klasifikasi Berita Politik Menggunakan Algoritma K-Nearest Neigbor," *Berkala Sainstek*, vol. VI, pp. 106-114, 2018.

[7] D. Sebastian, "Rancang Bangun Website Klasifikasi Untuk Pencarian Produk Pasar Online Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. III, pp. 417-432, 2017.

[8] M. Rivki and A. Bachtiar, "Implementasi Algoritma K-Nearest Neighbor Dalam Pengklasifikasian Follower Twitter yang Menggunakan Bahasa Indonesia," *Jurnal Sistem Informasi*, vol. XIII, pp. 31-37, 2017.

[9] A. D. M. Indonesia, "Digimind," [Online]. Available: https://digimind.id/10-kategori-produkterlaris-shopee. [Accessed 25 January 2021]