Available at https://jtsiskom.undip.ac.id (15 November 2019)
DOI:10.14710/jtsiskom.8.1.2020.54-58

Jurnal Teknologi dan Sistem Komputer, 8(1), 2020, 54-58

Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual

Comparison of distance measurement on k-nearest neighbour in textual data classification

Wahyono[1*)], I Nyoman Prayana Trisna[2)], Sarah Lintang Sariwening[2)], Muhammad Fajar[2)], Danur Wijayanto[2)]

[1)]Departemen Ilmu Komputer dan Elektronika, Fakultas MIPA, Universitas Gadjah Mada
Sekip Utara Bulaksumur, Kotak Pos 21, Senolowo, Sinduadi, Mlati, Sleman, DI Yogyakarta 55281

[2)]Program Magister Ilmu Komputer, Fakultas MIPA, Universitas Gadjah Mada
Sekip Utara Bulaksumur, Kotak Pos 21, Senolowo, Sinduadi, Mlati, Sleman, DI Yogyakarta 55281

Cara sitasi: W. Wahyono, I N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Perbandingan perhitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 1, pp. 54-58, 2020. doi: 10.14710/jtsiskom.8.1.2020.54-58, [Online].

Abstract – One algorithm to classify textual data in automatic organizing of documents application is KNN, by changing word representations into vectors. The distance calculation in the KNN algorithm becomes essential in measuring the closeness between data elements. This study compares four distance calculations commonly used in KNN, namely Euclidean, Chebyshev, Manhattan, and Minkowski. The dataset used data from Youtube Eminem's comments which contain 448 data. This study showed that Euclidian or Minkowski on the KNN algorithm achieved the best result compared to Chebycev and Manhattan. The best results on KNN are obtained when the K value is 3.

Keywords *- KNN; tekstual data; distance measurement; Euclidean; Chebyshev; Manhattan; Minkowski*

Abstrak - Salah satu algoritme yang dapat digunakan untuk mengklasifikasikan data tekstual dalam pengelolaan dokumen-dokumen secara otomatis adalah KNN, dengan mengubah representasi kata menjadi vektor. Penghitungan nilai jarak dalam algoritme KNN menjadi esensial dalam menentukan kedekatan antar elemen data. Penelitian ini membandingkan empat perhitungan jarak yang sering digunakan dalam KNN, yaitu Euclidean, Chebyshev, Manhattan, dan Minkowski. Dataset menggunakan data pada komentar Youtube Eminem yang berisi 448 data. Hasil penelitian ini menunjukkan bahwa jarak Euclidean dan Minkowski pada algoritme KNN pada data dengan representasi vektor dari kalimat sebagian besar menghasilkan akurasi terbaik dibandingkan Chebyshev maupun Manhattan. Hasil terbaik pada KNN diperoleh ketika K bernilai 3.

Kata kunci *- KNN; data teks; perhitungan distance; Euclidean; Chebyshev; Manhattan; Minkowski*

[*)] Penulis korespondensi (Wahyono)
Email: wahyo@ugm.ac.id

I. PENDAHULUAN

Text mining merupakan sebuah proses pengetahuan intensif dimana pengguna berinteraksi dan bekerja dengan sekumpulan dokumen dengan menggunakan beberapa alat analisis [1]. *Text mining* melakukan ekstraksi informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi dari suatu pola menarik. Suatu metode diperlukan untuk mengelola informasi dari sekumpulan dokumen teks yang jumlahnya sangat besar sehingga dapat mempermudah dalam pencarian informasi yang relevan dengan kebutuhan. Metode yang dapat mengorganisir dokumen teks secara otomatis di antaranya adalah klasifikasi.

Klasifikasi melakukan pengelompokan data dimana data yang digunakan tersebut mempunyai kelas label atau target. Algoritme untuk menyelesaikan masalah klasifikasi tersebut dimasukkan ke dalam *supervised learning* atau pembelajaran yang diawasi. Data label atau target ikut berperan sebagai supervisor atau guru yang mengawasi proses pembelajaran dalam mencapai tingkat akurasi atau presisi tertentu. Beberapa metode standar dapat digunakan untuk menyelesaikan masalah klasifikasi, yaitu di antaranya *backpropagation neural network*, *support vector classification* (SVC), *extreme learning machine* (ELM), *K-Nearest Neighbor* (KNN), dan Naïve Bayes.

Metode KNN melakukan klasifikasi terhadap objek berdasarkan data latih (*training*) yang menggunakan jarak terdekat atau kemiripan terhadap objek tersebut. Pada fase pembelajaran, algoritme ini melakukan penyimpanan vektor-vektor fitur dan klasifikasi dari data pembelajaran. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data pengujian (yang klasifikasinya tidak diketahui). Setelah jarak dari vektor yang baru terhadap seluruh vektor data pembelajaran dihitung dan sejumlah *K* buah yang paling dekat diambil, selanjutnya klasifikasi ditentukan dari titik-titik tersebut [1].

Perhitungan nilai jarak suatu data merupakan salah satu komponen esensial dari hasil proses algoritme

Submitted: 15 Juny 2019; Revised: 22 October 2019; Accepted: 5 November 2019; Published: 31 January 2020

KNN untuk mencocokkan antara data hasil pelatihan dengan data baru sebagai pembanding sehingga sangat mempengaruhi keakuratan kesamaan. Sebelum dilakukan pengelompokan data atau objek untuk proses deteksi, terlebih dahulu ditentukan ukuran jarak kedekatan antar elemen data. Untuk beragam aplikasi, beberapa metode pengukuran jarak digunakan untuk mengukur tingkat kesamaan (*similarity*) atau kemiripan suatu data, di antaranya menggunakan jarak Euclidean, Manhattan atau City Block Distance, Mahalanobis, Correlation, Angle-based, Minkowski, dan Squared Euclidean [2]-[8].

Cover dan Hart [2] melakukan pendekatan menggunakan tetangga K-terdekat. Tetangga terdekat dihitung berdasarkan nilai K. Pendekatan ini efektif jika ada data latih yang besar dan lemah pada data berukuran kecil. Bailey dan Jain [3] telah mempresentasikan KNN didasarkan pada bobot. Pendekatan ini untuk mengatasi kelemahan teknik sebelumnya, namun tidak terlalu signifikan. Menurut Prasath dkk. [4] performa KNN tergantung dari teknik penghitungan jarak yang digunakan, dibandingkan dengan mempertimbangkan ukuran K. Malkov dkk. [5] mengembangkan struktur KNN baru untuk data di ruang Euclidian berdasarkan graf dunia yang dapat dinavigasi dengan performansi yang lebih baik. Kirdat dan Patil [6] menggunakan jarak Chebyshev dan Minkowski untuk aplikasi CBIR menggunakan histogram warna. Isa dkk. [7] menyatakan jarak Minknowski menghasilkan akurasi lebih baik daripada Euclidean dalam klasifikasi sinyal EEG. Ali dkk. [8] menggunakan kombinasi jarak untuk data heterogen (numerik dan kategorikal) dibandingkan jarak Euclidean.

Kajian untuk membandingkan kinerja hasil KNN menggunakan teknik pengukuran jarak yang berbeda telah banyak dilakukan [9]-[14]. Chomboon dkk. [9] membandingkan kinerja jarak dengan delapan datasets dengan berbagai distribusi dan menyatakan bahwa City Block, Chebychev, Euclidean, Mahalanobis, Minkowski, dan Standardize Euclidean menghasilkan akurasi yang tinggi. Hu dkk. [10] menunjukkan bahwa fungsi pengukuran jarak Chi square memberikan hasil terbaik untuk tiga jenis dataset medis yang berbeda (tipe kategorikal, numerik, dan campuran), sedangkan pengukuran jarak Cosine, Euclidean, dan Minkowski menghasilkan performa paling buruk pada jenis kumpulan data campuran. Mulak dan Talhar [11] menyatakan bahwa penggunaan jarak Manhattan menghasilkan akurasi, sensitivitas, dan spesifisitas yang terbaik dibandingkan Chebyshev dan Euclidean pada dataset KDD. Sinwar dan Kaushik [12] menunjukkan bahwa Euclidean mempunyai performansi lebh baik daripada Manhattan dalam pengelompokan dengan kalkulasi centroid. Todeschini dkk. [13] menggunakan *meta-distances* untuk meningkatkan performansi dibandingkan jarak Euclidean. Singh dan Srivastava [14] menyatakan jarak Manhattan mempunyai performansi yang konsisten untuk aplikasi CBIR.

Kajian-kajian tersebut menunjukkan perbedaan kinerja metode pengukuran jarak terhadap performansi

```
COMMENT_ID:     z12zgrw5furdsn0sc233hfwavnznyhicq
AUTHOR:         kyeman13
DATE:
CONTENT:        Go check out my rapping video called Four Wheels please ♥
CLASS:          1

COMMENT_ID:     z12vxdzzds2kzzrzq04cdjc4ozq2szuyl5o
AUTHOR:         Damax
DATE:           2015-05-29T00:41:22.426000
CONTENT:        Almost 1 billion
CLASS:          0
```

Gambar 1. Contoh dua buah data komentar

KNN untuk beragam aplikasi yang berbeda. Penelitian ini bertujuan untuk mengkaji kinerja metode KNN terhadap 4 (empat) metode pengukuran jarak, yaitu jarak Euclidean, Chebyshev, Manhattan/City Block, dan Minkowski, untuk menganalisis tingkat akurasi yang dihasilkannya pada data dokumen tekstual. Setiap ukuran jarak dibandingkan dengan jumlah *K* yang berbeda, dengan *K* berjumlah ganjil dari 1 hingga 19. Kinerja KNN dengan masing-masing kombinasi jarak dan *K* dievaluasi berdasarkan akurasinya.

II. METODE PENELITIAN

Penelitian ini menggunakan data pada komentar Youtube Eminem yang berisi 448 data. Contoh data ditunjukkan pada Gambar 1. Dataset dapat diakses dari http://www.dt.fee.unicamp.br/~tiago//youtubespamcollection/. Penelitian ini hanya menggunakan kolom CONTENT sebagai variabel penentu dan CLASS sebagai target. CLASS 0 sebanyak 203 data menandakan bahwa komentar tersebut bukanlah *spam*, dan CLASS 1 sebanyak 245 data menandakan komentar *spam*.

Atribut CONTENT diubah representasinya menjadi vektor dengan menjadikan kata sebagai indeks dan frekuensi kata dalam komentar menjadi nilainya. Hal ini disebut sebagai *Bag-of-words* (BOW) [15]. Sebelum diubah ke dalam vektor, kata-kata yang bersifat *stopword* dihapuskan untuk mengurangi ukuran vektor. BOW yang terbentuk digunakan sebagai input classifier dan atribut CLASS sebagai output atau target. Alur penelitian ini ditunjukkan dalam Gambar 2. Parameter $f_{i,j}$ adalah kemunculan kata w_j di dalam komentar d_i, dengan i adalah jumlah komentar dan j jumlah kata.

Penelitian ini membandingkan empat ukuran jarak, yaitu Euclidean, Manhattan, Minkowski, dan Chebyshev. Masing-masing jarak ini adalah jarak antara dua buah titik yang memiliki atribut numerik. Jarak Euclidean dinyatakan dalam Persamaan 1, yang mendefinisikan total perbedaan kuadrat pada masing-masing atribut dua data. Bentuk lain perhitungan jarak adalah Jarak Chebyshev, yang hanya menggunakan selisih terbesar dari atribut-atribut dua buah data sebagai jarak dan dinyatakan dalam Persamaan 2.

$$D(d_x, d_y) = \sqrt{(f_{x,1} - f_{y,1})^2 + \ldots + (f_{x,j} - f_{y,j})^2} \quad (1)$$

$$D(d_x, d_y) = max\left(|f_{x,1} - f_{y,1}|, \ldots, |f_{x,j} - f_{y,j}|\right) \quad (2)$$

Bentuk perhitungan jarak yang lain dalam penelitian ini adalah jarak Manhattan atau City Block Distance.

Seperti Euclidean, jarak Manhattan mempertimbangkan selisih atribut pada dua data dengan menggunakan selisih absolut. Perhitungan jarak Manhattan dinyatakan dalam Persamaan 3.

$$D(d_x, d_y) = |f_{x,1} - f_{y,1}| + \ldots + |f_{x,j} - f_{y,j}| \quad (3)$$

Penelitian ini juga mengimplementasikan jarak Minkowski. Jarak Minkowski mempunyai bentuk perhitungan jarak lain yang mirip seperti Euclidean dan Manhattan, namun pangkat dan akar pangkat *p* yang digunakan berkisar dari 1 hingga 2. Perhitungan jarak Minkowski dinyatakan pada Persamaan 4. Nilai *p* yang digunakan dalam penelitian ini adalah 1,5.

$$D(d_x, d_y) = \sqrt[p]{(f_{x,1} - f_{y,1})^p + \ldots + (f_{x,j} - f_{y,j})^p} \quad (4)$$

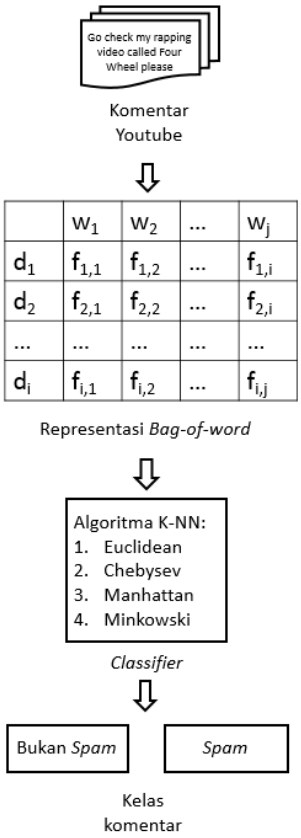
Penelitian ini membandingkan 4 buah ukuran jarak dengan jumlah *K* yang berbeda. Jumlah *K* yang ditentukan berjumlah ganjil, dari 1 hingga 19, sehingga terdapat 40 hasil yang dianalisis. Masing-masing kombinasi jarak dan *K* dievaluasi dengan menghitung akurasinya. Akurasi dihitung dengan menjumlahkan data yang tepat diprediksi, apakah komentar tersebut *spam* atau bukan, terhadap total data keseluruhan. Perhitungan akurasi dinyatakan dalam Persamaan 5. Akurasi dihitung dengan menggunakan data pelatihan sebanyak 90 % dan data pengujian sebanyak 10 %. Pengujian dilakukan dengan *cross validation* untuk mendapatkan hasil secara menyeluruh.

$$akurasi = \frac{\sum data\ diprediksi\ tepat}{\sum data} \quad (5)$$

III. HASIL DAN PEMBAHASAN

Hasil penelitian berupa nilai akurasi pada setiap algoritme jarak yang digunakan dengan input berupa vektor dokumen dalam bentuk BOW dan output berupa *spam* atau bukan. Hasil akurasi didapatkan dari variasi nilai *K* yang digunakan. BOW yang terbentuk adalah berupa nilai skalar yang berdimensi 1433, yaitu terdapat 1433 kata unik di dalam keseluruhan dokumen setelah dilakukan *stopword removal*. Sebagai contoh vektorisasi, komentar "Almost hit billions" ditransformasi menjadi nilai skalar sebesar 1433 yang semua indeksnya bernilai 0, kecuali pada indeks 186 dan 599 karena 186 adalah indeks untuk kata "billions" dan 599 adalah indeks untuk kata "hit". Dengan seluruh komentar yang digunakan, penelitian ini menggunakan vektor berukuran 448 × 1433.

Hasil dari perhitungan jarak Euclidean, Chebyshev, Manhattan, dan Minkowski terhadap akurasi KNN ditunjukkan pada Tabel 1. Hasil penelitian menunjukkan bahwa semakin banyak nilai *K* menyebabkan penurunan akurasi, walaupun ada kondisi dimana nilai K tertentu akurasi yang meningkat. Namun, hal ini tidak berlaku untuk jarak Chebyshev. Akurasi terendah diperoleh dari jarak Chebyshev, sedangkan akurasi tertinggi



Gambar 2. Alur penelitian keseluruhan

Tabel 1 Akurasi tiap model percobaan terhadap nilai K

K	Euclidean	Chebyshev	Manhattan	Minkowski
1	0,8504	0,6585	0,8548	0,8504
3	0,8550	0,6187	0,8505	0,8550
5	0,8372	0,6187	0,8325	0,8372
7	0,8372	0,6096	0,8259	0,8372
9	0,8216	0,6183	0,8079	0,8216
11	0,8128	0,6119	0,7968	0,8128
13	0,8041	0,6359	0,7814	0,8041
15	0,8041	0,6359	0,7814	0,8041
17	0,7971	0,6047	0,7724	0,7971
19	0,7972	0,5960	0,7679	0,7972

didapatkan dari jarak Euclidean dan Minkowski yang mempunyai nilai sama setiap *K*.

Pada *K* = 1, jarak Manhattan cocok pada dimensi data vektor seperti data pada penelitian ini, yaitu berdimensi 1433. Hal ini disebabkan karena tidak ada bentuk polinomial di dalam perhitungan jarak Manhattan dan hanya menggunakan selisih absolut tiap elemen dalam vektor. Mempertimbangkan hanya satu *neighbour*, perhitungan selisih atribut tanpa bentuk polinomial menyebabkan nilai akurasi yang lebih baik, walaupun tidak berbeda jauh dengan bentuk kuadratik seperti jarak Euclidean. Hal ini selaras dengan [11] yang menunjukkan bahwa penggunaan jarak Manhattan sedikit lebih baik dibandingkan jarak Euclidean pada nilai K yang kecil.

Jarak Euclidean dan Minkowski dengan p = 1,5 memiliki akurasi yang sama pada tiap penambahan jumlah K. Hu dkk. [10] menunjukkan hal yang sama, di mana data dengan bentuk berbeda menghasilkan akurasi yang sama apabila menggunakan jarak Euclidean maupun Minkowski. Hal ini menandakan bahwa penggunaan nilai p = 2 maupun p = 1,5 tidak mempengaruhi kinerja KNN pada dataset penelitian ini maupun dataset [10]. Kesesuaian antara dua kajian ini adalah penggunaan data yang memiliki kategori biner atau sedikit.

Jarak Euclidean maupun jarak Minkowski menghasilkan akurasi terbaik pada sebagian besar jumlah K, walaupun selisih hasilnya tidak terlampau jauh dengan jarak Manhattan seperti [12]. Hal ini menandakan bahwa melebih-lebihkan perbedaan pada 2 buah data dengan melakukan pemangkatan terhadap selisih atribut 2 data tersebut menghasilkan jarak yang lebih baik untuk klasifikasi biner dibandingkan hanya dengan mengandalkan selisih absolut 2 buah data. Perbedaan yang tidak terlampau jauh antara Euclidean, Minkowski, dan Manhattan terjadi karena ketiga perhitungan jarak ini menggunakan seluruh atribut dalam mencari jarak dua buah data. Hal ini berbeda dengan jarak Chebyshev.

Jarak Chebyshev memiliki hasil terburuk berdasarkan akurasi dari tiap K. Hal ini berbeda dengan [9] yang memperlihatkan bahwa jarak Chebyshev memiliki akurasi yang tidak jauh berbeda dengan Manhattan, Euclidean, maupun Minkowski. Pada [9], dataset yang digunakan adalah dataset sintesis yang hanya memiliki 3 atribut, sedangkan pada penelitian ini terbentuk 1433 atribut. Oleh karena jarak Chebyshev hanya mempertimbangkan satu atribut dengan selisih maksimal pada dua buah data, maka jarak Chebyshev handal pada data berdimensi rendah. Sebaliknya, jika dimensi data sangat besar seperti BOW, maka jarak Chebyshev hanya dihitung pada satu atribut tanpa menghitung atribut lain.

Pada data teks, jarak Chebyshev hanya menghitung jarak antara kalimat d_1 dan d_2 berdasarkan kata yang paling sering muncul di kalimat d_1 namun tidak muncul sama sekali di kalimat d_2 atau sebaliknya. Dari hasil akurasi ini, dapat dilihat bahwa klasifikasi komentar *spam* atau bukan tidak dapat hanya dengan melihat perbedaan satu kata pada dua buah komentar.

Secara umum, penambahan jumlah K pada tiap perhitungan jarak menghasilkan akurasi yang menurun, walaupun hasil terbaik diperoleh saat K = 3. Hal ini menandakan penambahan jumlah K membuat misklasifikasi. Hal ini disebabkan karena kalimat komentar yang memiliki kelas yang sama cenderung dekat dengan sedikit anggota dari kelas tersebut. Ketika nilai K dinaikkan, maka kalimat komentar tersebut justru lebih banyak dekat ke kelas lainnya. Dengan melihat penurunan akurasi pada Tabel 1, dapat dinyatakan bahwa suatu komentar tidak hanya dekat ke kelasnya sendiri, namun juga didekati oleh kelas lainnya. Perhitungan jarak dengan Manhattan merupakan perhitungan yang paling sensitif dengan perubahan jumlah K seperti [11], [14].

IV. KESIMPULAN

Jarak Euclidean maupun Minkowski dengan p sebesar 1,5 menghasilkan akurasi terbaik pada sebagian besar ukuran K, sedangkan jarak Chebyshev menjadi perhitungan jarak terburuk berdasarkan akurasinya. Penambahan jumlah K pada tiap perhitungan jarak menurunkan akurasi *classifier* sebab jumlah K yang makin banyak membuat lebih banyak data yang tidak memiliki kelas yang sama dibandingkan kelas yang tepat. Jumlah K = 3 adalah jumlah K paling efektif.

DAFTAR PUSTAKA

[1] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[2] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967. doi: 10.1109/TIT.1967.1053964

[3] T. Bailey and A. K. Jain, “A note on distance-weighted k-nearest neighbor rules,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 4, pp. 311-313, 1978. doi: 10.1109/TSMC.1978.4309958

[4] V. B. Prasath et al., “Distance and similarity measures effect on the performance of k-nearest neighbor classifier - a review,“ *arXiv:1708.04321 v3 [cs.LG]*, 2019. doi: 10.1089/big.2018.0175

[5] Y. Malkov, A. Ponomarenko, A. Logvinov, and V. Krylov, "Approximate nearest neighbor algorithm based on navigable small world graphs," *Information Systems*, vol. 45, pp. 61–68, 2014. doi: 10.1016/j.is.2013.10.006

[6] T. Kirdat and V. V. Patil, “Application of Chebyshev distance and Minkowski distance to CBIR using color histogram,” *International Journal of Innovative Research in Technology (IJIRT)*, vol. 2, no. 9, pp. 28-31, 2016.

[7] N. E. Md Isa, A. Amir, M. Z. Ilyas, and M. S. Razalli, “The performance analysis of k-nearest neighbors (K-NN) algorithm for motor imagery classification based on EEG signal, ” in *2017 International Conference on Emerging Electronic Solutions for IoT*, Penang, Malaysia, Oct. 2017, pp. 1-6. doi: 10.1051/matecconf/201714001024

[8] N. Ali, D. Neagu, and P. Trundle, “Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets,” *SN Applied Sciences*, vol. 1, no. 1559, 2019. doi: 10.1007/s42452-019-1356-9

[9] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, and N. Kerdprasop, “An empirical study of distance metrics for k-nearest neighbor algorithm,” in *the 3rd International Conference on*

Industrial Application Engineering 2015, Kitakyushu, Japan, Mar. 2015, pp. 1-6. doi: 10.12792/iciae2015.051

[10] L. Y. Hu, M. W. Huang, S. W. Ke, and C. F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, pp. 1-9, 2016. doi: 10.1186/s40064-016-2941-7

[11] P. Mulak and N. Talhar, "Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset," *International Journal of Science and Research*, vol. 4, no. 7, pp.2101-2104, 2015.

[12] D. Sinwar and R. Kaushik, "Study of Euclidean and Manhattan distance metrics using simple k-means clustering," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 2, no. 5, pp. 270-274, 2014.

[13] R. Todeschini, D. Ballabio, V. Consonni, and F. Grisoni, "A new concept of higher-order similarity and the role of distance/similarity measures in local classification methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 157, pp. 50-57, 2016. doi: 10.1016/j.chemolab.2016.06.013

[14] Y. Shikhar, V. P. Singh, and R. Srivastava, "Comparative analysis of distance metrics for designing an effective content-based image retrieval system using colour and texture features," *International Journal of Image, Graphics, and Signal Processing*, vol. 12, no. 7, pp. 58-65, 2017. doi: 10.5815/ijigsp.2017.12.07

[15] P. Koniusz, F. Yan, P. H. Gosselin, and K. Mikolajczyk, "Higher-order occurrence pooling for bags-of-words: visual concept detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 313–326, 2017. doi: 10.1109/TPAMI.2016.2545667