IMPLEMENTASI COSINE SIMILARITY UNTUK PENINGKATAN AKURASI PENGUKURAN KESAMAAN DOKUMEN PADA KLASIFIKASI DOKUMEN BERITA DENGAN K NEAREST NEIGHBOUR

Firdaus[1], Pasnur[2], Wabdillah[3]
RSUP Dr. Wahidin Sudirohusodo, Makassar, Sulawesi Selatan, Indonesia[1]
STMIK AKBA, Makassar, Sulawesi Selatan, Indonesia[2,3]
Email : firdaus.wahidin@gmail.com[1], pasnur@akba.ac.id[2], wabdillah@akba.ac.id[3]

ABSTRAK

Klasifikasi dokumen berita secara otomatis menggunakan komputer diusulkan agar lebih efisien dalam memproses dokumen dalam jumlah banyak. Metode klasifikasi *K-Nearest Neighbour* yang menggunakan *Euclidean Distance* sebagai metode pengukuran kesamaan dokumen kurang akurat karena dipengaruhi oleh panjang dokumen. Dokumen yang mirip tetapi memiliki panjang dokumen yang berbeda mungkin memiliki nilai jarak yang tinggi. Tujuan penelitian ini adalah mengimplementasikan *Cosine Similarity* untuk meningkatkan akurasi pengukuran kesamaan dokumen pada klasifikasi dokumen berita dengan metode *K-Nearest Neighbor*. Pada penelitian ini diusulkan *Cosine Similarity* sebagai metode pengukuran kesamaan dokumen berita. *Cosine Similarity* menghitung kesamaan antar dua buah dokumen berdasarkan besar sudut cosinus. Hasil penelitian menunjukkan bahwa implementasi *Cosine Similarity* dapat meningkatkan akurasi pengukuran kesamaan dokumen pada klasifikasi dokumen berita dengan metode *K-Nearest Neighbour*. Rata-rata akurasi metode *K-Nearest Neighbour* dengan *Cosine Similarity* adalah 98,12%, sedangkan akurasi metode *K-Nearest Neighbour* dengan *Euclidean Distance* adalah 56,51%.

Kata Kunci: *Cosine Similarity, Euclidean Distance, K-Nearest Neighbour, Text Classification*

ABSTRACT

News documents automatic classification by using computer was proposed for more efficient process to handle a large of documents. K-Nearest Neighbour classification with Euclidean Distance as document similarity measurement is less accurate because affected by document size. Similar documents with different size may have higher distance. This study aims to use Cosine Similarity to increase documents similarity measurements in new documents classification with K-Nearest Neighbour. This study proposed Cosine Similarity as news documents similarity measurements. Cosine Similarity calculates similarity of two documents according to a value of cosine degree. The experimental result shows that the implementation of Cosine Similarity can increase the accuracy of document similarity measurements in news documents classification with K-Nearest Neighbour. The average accuracy of K-Nearest Neighbour with Cosine Similarity and Euclidean Distance is 98,12% and 56,51%..

Keywords: *Cosine Similarity, Euclidean Distance, K-Nearest Neighbour, Text Classification*

1. PENDAHULUAN

Information retrieval memiliki banyak aplikasi dalam kehidupan sehari-hari. Beberapa di antaranya adalah pencarian dan perangkingan dokumen teks tradisional maupun dengan melibatkan semantik (Chouni, Erritali, Madani, & Ezzikouri, 2019; Wahib, Pasnur, Santika, & Arifin,

2015), pengelompokan dokumen teks (Jiang, Pang, Wu, & Kuang, 2012), dan peringkasan dokumen otomatis (Pasnur, Santika, & Syaifuddin, 2014). Pengelompokan atau klasifikasi dokumen teks (dokumen berita, surat elektronik, dokumen ilmuah) termasuk topik yang sangat diminati oleh para peneliti.

Klasifikasi dokumen berita secara manual tidak efisien. Pengelola dokumen berita pada situs-situs berita online akan kewalahan apabila mereka harus mengelompokkan dokumen-dokumen berita secara manual. Teknik klasifikasi manual memerlukan waktu yang lama, terutama jika jumlah dokumen berita sangat banyak. Pada kondisi seperti ini, maka dibutuhkan sebuah sistem yang mampu melakukan klasifikasi dokumen berita secara otomatis seperti pada penelitian (Jiang et al., 2012).

Berbagai metode klasifikasi dokumen teks telah diusulkan oleh para peneliti, seperti *Decision Tree (DT)* (Apte, Damerau, Sholom, & Weiss, 1994), *K Nearest Neighbour (KNN)* (Guo, Wang, & Bell, 2004), *Naive Bayes (NB)* (Frank & Bouckaert, 2006), *Neural Network (Nnet)* (Ruiz & Srinivasan, 2002), *Support Vector Machine (SVM)* (Chen & Hsieh, 2006), dan *Centroid-based Approaches* (Tan, 2008). Metode-metode yang telah diusulkan masing-masing memiliki kelebihan dan kekurangan. Metode-metode tersebut juga dapat diimplementasikan pada klasifikasi dokumen berita, seperti halnya jenis dokumen teks yang lain.

Metode KNN merupakan metode yang sederhana tetapi efektif dalam melakukan klasifikasi dokumen teks (Jiang et al., 2012). Akan tetapi, penggunaan KNN dalam klasifikasi dokumen teks juga dipengaruhi oleh metode pengukuran kesamaan dokumen yang digunakan. Metode pengukuran similaritas dan jarak seperti *euclidean distance* dan *jaccard coefficent* sangat dipengaruhi oleh panjang dokumen. Pada pengukuran *euclidean distance*, dokumen yang mirip tetapi memiliki panjang dokumen yang berbeda akan sangat mungkin memiliki nilai jarak yang tinggi. Hal tersebut akan menurunkan akurasi klasifikasi.

Pada penelitian ini akan diusulkan pengukuran kesamaan dokumen menggunakan *cosine similarity* pada klasifikasi dokumen berita menggunakan KNN. Pengukuran *cosine similarity* merupakan metode yang digunakan untuk menghitung tingkat kesamaan antar dua buah objek berdasarkan besar sudut *cosinus*. Nilai sudut *cosinus* antara dua vektor menentukan kesamaan dua buah objek yang dibandingkan dimana nilai terkecil adalah 0 dan nilai terbesar adalah 1 (Nurdiana, Jumadi, & Nursantika, 2016).

2. LANDASAN TEORI

Klasifikasi dokumen berita merupakan salah satu contoh aplikasi dari *information retrieval*. Klasifikasi dokumen berita juga berkaitan dengan *text mining*. Tahapan dalam klasifikasi dokumen berita secara umum meliputi persiapan *dataset*, *preprocessing*, pembobotan *term*, pengukuran similaritas, serta penentuan hasil klasifikasi.

Pada klasifikasi dokumen berita, setiap *term* dari seluruh dokumen akan diberikan bobot. *Term* yang sering muncul pada sebuah dokumen memiliki bobot informasi yang lebih tinggi. Nilai bobot frekuensi kemuculan *term* pada sebuah dokumen disebut dengan *term frequency (TF)*. *Term* yang muncul hanya pada sedikit dokumen juga memiliki bobot yang lebih tinggi. Nilai bobot jumlah dokumen yang memiliki sebuah *term* tertentu disebut *inverse document frequency (IDF)* (Wahib et al., 2015).

Bobot informasi *term* yang digunakan pada penelitian ini adalah TF.IDF dan merupakan nilai vektor dari dokumen yang akan diproses (Wahib et al., 2015). Bobot ini merupakan perkalian antara nilai TF dan IDF seperti yang ditunjukkan pada Persamaan (1).

$$w_{ij} = tf_{ij}\, x\, idf \quad (1)$$

Nilai w_{ij} merupakan bobot TF.IDF, sedangkan nilai tf_{ij} dan idf masing-masing menunjukkan bobot TF dan IDF. Nilai TF dapat dilakukan dengan menghitung frekuensi kemunculan *term* pada sebuah dokumen tertentu. Sedangkan nilai IDF dapat dihitung dengan menggunakan Persamaan (2).

$$idf = 1 + log\left(\frac{N}{df_j}\right) \quad (2)$$

Pada Persamaan (2), *N* merupakan jumlah seluruh dokumen yang ada pada dataset, sedangan df_j merupakan jumlah dokumen yang memiliki *term j*.

Pada pengukuran similaritas dengan *cosine similarity*, nilai TF.IDF merupakan nilai setiap dimensi dokumen yang akan dibandingkan. Nilai *cosine similarity* antara 2 buah vektor dokumen A dan B, dapat dihitung menggunakan Persamaan (3).

$$Sim_{A,B} = \frac{AxB}{|A||B|} = \frac{\sum_{i=1}^{n} A_i x B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\, x \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

Pada Persamaan (3), AxB merupakan *dot product* antara dokumen A dan B, sedangkan |A|.|B| merupakan perkalian panjang vektor antara dokumen A dan B. *Dot product* dilakukan dengan melakukan perkalian biasa antara nilai TF.IDF dokumen A dengan dokumen B untuk setiap dimensi. Sedangkan panjang vektor A dan B, dapat diketahui dengan cara menghitung akar kuadrat dari penjumlahan pangkat dua masing-masing nilai TF.IDF pada setiap dokumen A dan B.

Hasil pengukuran similaritas dokumen menggunakan *cosine similarity* akan digunakan oleh metode klasifikasi KNN untuk menentukan hasil klasifikasi dokumen berita. Metode KNN menentukan hasil klasifikasi berdasarkan *class* dominan dari sejumlah *k* dokumen terdekat dengan dokumen uji. Pada proses pengujian, diupayakan untuk menghindari penggunaan nilai *k* yang menyebabkan jumlah *class* dominan yang seimbang.

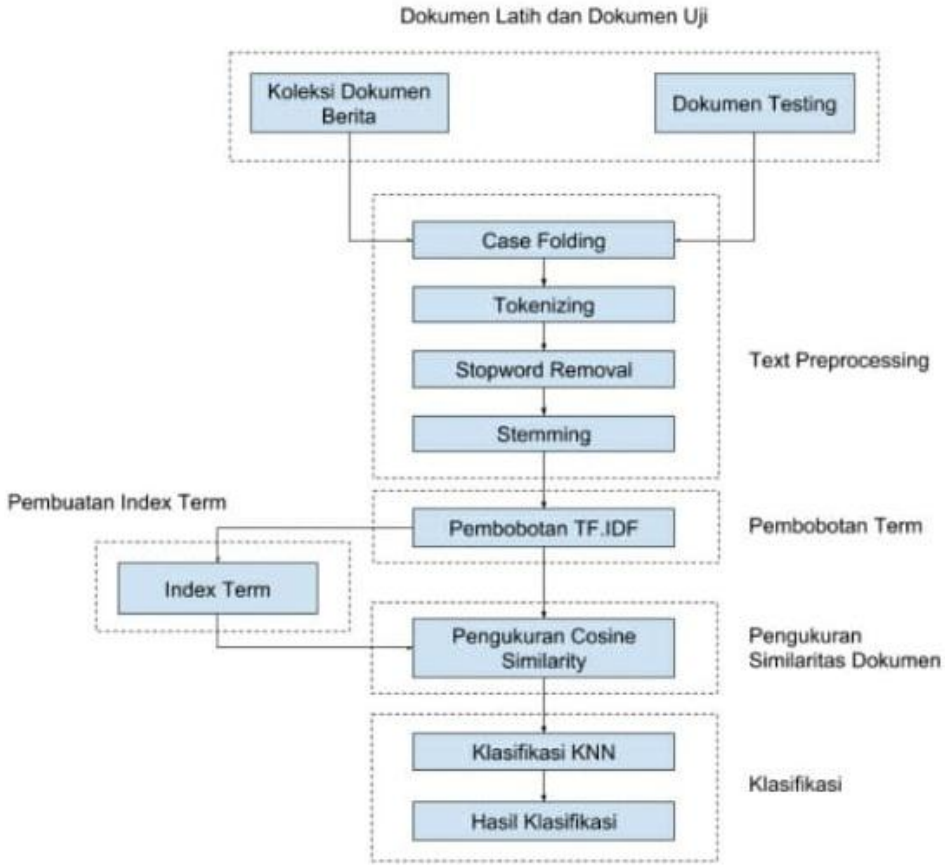
3. METODE PENELITIAN

Pada penelitian ini, diusulkan model sistem seperti pada Gambar (1). Tahapan yang digunakan meliputi persiapan dokumen berita, *preprocessing*, pembobotan *term*, pengukuran similaritas dokumen berita, dan klasifikasi dokumen berita.

Dokumen berita yang digunakan pada penelitian ini diambil dari situs berita *online* kompas.com. Jumlah dokumen berita yang digunakan sebanyak 150, yang terbagi ke dalam tiga kategori, yaitu berita tekno, kesehatan, dan edukasi. Masing-masing kategori memiliki jumlah dokumen sebanyak 50.

Setiap dokumen latih dan uji akan melalui tahapan *preprocessing.* Pada tahap ini setiap karakter pada dokumen akan diubah menjadi huruf kecil *(case folding).* Hal ini untuk mendapatkan bentuk huruf yang seragam. Selanjutnya akan dilakukan pemecahan dokumen menjadi bentuk kecil yang disebut *token (tokenizing).* Pada penelitian ini, *token* yang digunakan berbentuk kata/*term.* Hasil *tokenizing* akan diseleksi melalui tahap *stopword removal.* Pada tahap ini, akan dilakukan penghapusan kata-kata yang termasuk kategori *stopword.* Kata-kata tersebut memiliki frekuensi kemunculan tinggi pada dokumen, tetapi memiliki bobot informasi yang sangat rendah. Pada setiap kata yang telah diseleksi, akan dilakukan perubahan morfologi menjadi bentuk kata dasar, dengan membuang awalan dan akhiran pada kata tersebut.

Setelah melalui tahap *preprocessing*, maka akan dilakukan beberapa tahapan inti, meliputi pembobotan *term* dengan menggunakan TF.IDF, pembuatan *index*, pengukuran similaritas dokumen dengan menggunakan *cosine similarity*, dan penentuan klasifikasi dokumen berita menggunakan *KNN*. Pada pengujian KNN akan digunakan berbagai nilai k yaitu k={1, 5, 10, 15, 20, 25, 30}.



Gambar 1. Model Sistem Usulan

4. HASIL DAN PEMBAHASAN

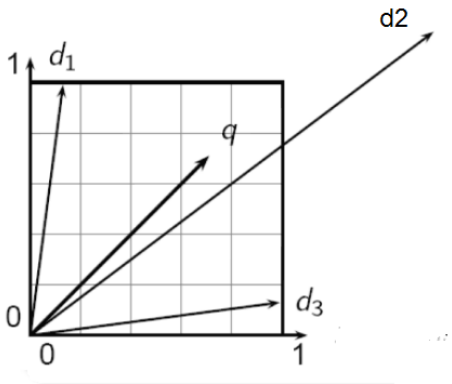
Pengujian dilakukan dengan membandingkan akurasi klasifikasi KNN dengan pengukuran similatitas *euclidean distance* dan *cosine similarity*. Rekapitulasi nilai akurasi hasil pengujian ditunjukkan pada Tabel (1). Nilai akurasi dinyatakan dalam satuan persentase yang diuji pada berbagai nilai *k* yang berbeda. Nilai akurasi tertinggi dengan *cosine similarity* didapatkan pada nilai *k=10* yaitu sebesar 99,06%. Sedangkan nilai akurasi tertinggi untuk *euclidean distance* didapatkan pada nilai *k=1* yaitu sebesar 60,05%.

Tabel 1. Akurasi Pengujian Klasifikasi KNN dengan *Euclidian Distance* dan *Cosine Similarity*

k	Akurasi (%)	
	Euclidean Distance	*Cosine Similarity*
1	60,05	98,59
5	56,60	98,59
10	55,70	99,06
15	55,70	98,12
20	55,70	97,65
25	56,13	97,65
30	55,70	97,18

Berdasarkan rekapitulasi hasil pengujian, didapatkan hasil bahwa pengukuran similaritas dokumen dengan menggunakan *cosine similarity* akan menghasilkan akurasi klasifikasi *KNN* yang lebih tinggi dibandingkan dengan *euclidean distance*. Apabila dilakukan perhitungan nilai akurasi rata-rata dari seluruh nilai *k* yang diuji, maka didapatkan nilai akurasi rata-rata klasifikasi *KNN* dengan *cosine similarity* adalah 98,12% dan *euclidean distance* 56,51%.

Nilai akurasi klasifikasi *KNN* yang menggunakan *cosine similarity* lebih tinggi dari pada *euclidean distance*, disebabkan oleh ketelitian pengukuran sudut vektor dokumen. Pada Gambar (2) ditunjukkan perbandingan pengkuran kemiripan dokumen kedua metode tersebut.



Gambar 2. Perbandingan Pengukuran Kemiripan Dokumen dengan *Cosine Similarity* dan *Euclidean Distance*

Pada Gambar (2), terlihat bahwa dokumen *q* dan dokumen *d2* memiliki kemiripan yang tinggi. Kedua dokumen tersebut memiliki kesamaan kata-kata yang tinggi, sehingga vektor kedua dokumen memiliki sudut yang kecil. Dokumen *q* memiliki panjang dokumen yang lebih kecil dari pada dokumen *d2*.

Pengukuran dengan menggunakan *euclidean distance* pada Gambar (2), akan mendapatkan hasil pengukuran jaran *q-d1* lebih dekat dibandingkan dengan jarak *q-d2*. Hal ini menunjukkan bahwa dokumen *d1* lebih relevan dengan *q* dibandingkan dengan *d2*. Hasil pengukuran ini tidak sesuai dengan fakta bahwa dokumen *d2* lebih relevan dengan *q* dibandingkan dengan *d1*.

Teori dan pembuktian empiris melalui eksperimen menunjukkan bahwa dalam pengukuran similaritas dokumen teks dibutuhkan pengkuran sudut vektor dokumen untuk mendapatkan hasil yang lebih akurat. Pengukuran jarak seperti yang digunakan pada *euclidean distance* kurang efektif pada klasifikasi dokumen teks karena akan dipengaruhi oleh panjang dokumen.

Pengukuran similaritas yang lebih efektif untuk klasifikasi dokumen teks adalah menggunakan pengukuran sudut *cosinus* seperti pada *cosine similarity*. Pada pengukuran *cosine similarity*, kesamaan dokumen tidak terlalu dipengaruhi oleh panjang dokumen. Nilai kesamaan dokumen berdasarkan besar sudut vektor kedua dokumen yang dibandingkan yang dihitung menggunakan nilai *cosinus.* Pengukuran dengan nilai *cosine similarity* lebih tinggi menunjukkan nilai similaritas lebih tinggi dibandingkan dengan nilai *cosine similarity* lebih rendah.

5. SIMPULAN DAN SARAN

Berisi ringkasan hasil penelitian dan rekomendasi penulis terkait pengembangan ilmu, teknologi maupun inovasi di bidang komunikasi, informatika dan media massa.

DAFTAR PUSTAKA

Apte, C., Damerau, F., Sholom, M., & Weiss. (1994). Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems (TOIS)*, *12*(3), 233–251.

Chen, R. C., & Hsieh, C. H. (2006). Web Page Classification Based on a Support Vector Machine using a Weighted Vote Schema.

Expert Systems with Applications, *31*(2), 427–435.

Chouni, Y., Erritali, M., Madani, Y., & Ezzikouri, H. (2019). Information Retrieval System based Semantique and Big Data. *Procedia Computer Science*, *151*, 1108–1113. https://doi.org/10.1016/j.procs.2019.04.157

Frank, E., & Bouckaert, R. (2006). Naive Bayes for Text Classification with Unbalanced Classes. *Knowledge Discovery in Databases*, 503–510.

Guo, G., Wang, H., & Bell, D. (2004). KNN Model-Based Approach and Its Application in Text Categorization. *Computational Linguistics and Intelligent Text Processing, LNCS*, 559–570.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An Improved K-Nearest-Neighbor Algorithm for Text Categorization. *Expert Systems with Applications*, *39*(1), 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

Lu, Y., He, H., Zhao, H., Meng, W., & Yu, C. (2013). Annotating Search Results from Web Databases. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 514–527. https://doi.org/10.1109/TKDE.2011.175

Nurdiana, O., Jumadi, J., & Nursantika, D. (2016). Perbandingan Metode Cosine Similarity dengan Metode Jaccard Similarity pada Aplikasi Pencarian Terjemah Al-Qur'an dalam Bahasa Indonesia. *Jurnal Online Informatika (JOIN)*, *1*(1), 59–63. Retrieved from https://doi.org/10.1177/01945998114098662

Pasnur, P., Santika, P. P., & Syaifuddin, G. N. (2014). Semantic Clustering dan Pemilihan Kalimat Representatif untuk Peringkasan Multi Dokumen. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *1*(2), 91–97. Retrieved from http://jtiik.ub.ac.id/index.php/jtiik/article/view/117

Ruiz, M., & Srinivasan, P. (2002). Hierarchical Text Categorization using Neural Networks. *Information Retrieval*, *5*(1), 87–118.

Tan, S. (2008). An Improved Centroid Classifier for Text Categorization. *Expert Systems with Applications*, *35*(1), 279–285.

Wahib, A., Pasnur, P., Santika, P. P., & Arifin, A. Z. (2015). Perangkingan Dokumen Berbahasa Arab Menggunakan Latent Semantic Indexing. *Jurnal Buana Informatika*, *6*(2), 83–92. Retrieved from https://ojs.uajy.ac.id/index.php/jbi/article/view/411

Zhang, J., Feng, S., Li, D., Gao, Y., Chen, Z., & Yuan, Y. (2017). Image Retrieval Using The Extended Salient Region. *Information Sciences*, *399*, 1339–1351. https://doi.org/10.1016/j.ins.2017.03.005