JISKa, Vol. 4, No. 3, JANUARI, 2020, Pp. 185 – 192
ISSN : 2527 – 5836 (print) | 2528 – 0074 (online)

Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine

Muhammad Habibi[(1)], Puji Winar Cahyo[(2)]
Program Studi Informatika
Universitas Jenderal Achmad Yani Yogyakarta
e-mail : muhammadhabibi17@gmail.com[(1)], pwcahyo@gmail.com[(2)]

Abstract

One of the problems related to journal publishing is the process of categorizing entry into journals according to the field of science. A large number of journal documents included in a journal editorial makes it difficult to categorize so that the process of plotting to reviewers requires a long process. The review process in a journal must be done planning according to the expertise of the reviewer, to produce a quality journal. This study aims to create a classification model that can classify journals automatically using the Cosine Similarity algorithm and Support Vector Machine in the classification process and using the TF-IDF weighting method. The object of this research is abstract in scientific journals. The journals will be classified according to the reviewer's field of expertise. Based on the experimental results, the Support Vector Machine method produces better performance accuracy than the Cosine Similarity method. The results of the calculation of the value of precision, recall, and f-score are known that the Support Vector Machine method produces better amounts, in line with the accuracy value.

Keywords : Text Mining, Cosine Similarity, Classification, Journal, Support Vector Machine

Abstrak

Salah satu masalah yang berkaitan dengan penerbitan jurnal yaitu proses pengkategorian jurnal masuk sesuai dengan bidang ilmu. Banyaknya jumlah dokumen jurnal yang masuk dalam suatu editorial jurnal membuatnya sulit untuk dilakukan pengkategorian sehingga proses plotting kepada *reviewer* membutuhkan proses yang lebih lama. Proses *review* pada suatu jurnal harus dilakukan plotting menyesuaikan dengan bidang keahlian dari *reviewer*, sehingga menghasilkan jurnal yang berkualitas. Penelitian ini bertujuan untuk membuat model klasifikasi yang dapat mengklasifikasikan jurnal secara otomatis menggunakan algoritma *Cosine Similarity* dan *Support Vector Machine* dalam proses pengklasifikasiannya dan menggunakan metode pembobotan TF-IDF. Objek penelitian ini adalah *abstract* pada jurnal ilmiah. Jurnal akan diklasifikasikan sesuai dengan rumpun ilmu bidang keahlian dari *reviewer*. Berdasarkan hasil eksperimen, metode *Support Vector Machine* menghasilkan akurasi performansi yang lebih baik dari pada metode *Cosine Similarity*. Hasil perhitungan nilai *precision*, *recall*, dan *f-score* diketahui bahwa metode *Support Vector Machine* menghasilkan nilai yang lebih baik, sejalan dengan nilai akurasi.

Kata Kunci : *Text Mining*, *Cosine Similarity*, Klasifikasi, Jurnal, *Support Vector Machine*

1. PENDAHULUAN

Perkembangan teknologi informasi membawa dampak yang sangat signifikan pada dunia Pendidikan. Salah satunya adalah melimpahnya informasi yang dapat diakses sebagai referensi dalam Pendidikan. Penyebaran jurnal atau artikel ilmiah sebagai bahan pendukung penelitian semakin meningkat. Dalam prosesnya, penerbitan jurnal memiliki beberapa tahapan mulai dari submission jurnal sampai jurnal tersebut terbit. Banyaknya jumlah dokumen jurnal yang masuk dalam suatu editorial jurnal membuatnya sulit untuk dilakukan pengkategorian sehingga proses *plotting* kepada *reviewer* membutuhkan proses yang lebih lama. Proses *review*

pada suatu jurnal harus dilakukan *plotting* menyesuaikan dengan bidang keahlian dari *reviewer*, sehingga menghasilkan jurnal yang berkualitas.

Untuk mempermudah mengkategorikan jurnal, diperlukan teknik pemrosesan teks yang dapat mengkategorikan sejumlah besar dokumen teks sesuai dengan tipenya, sehingga informasi yang tersedia dapat diakses dengan benar dan mudah diakses sesuai dengan kebutuhan pengguna. Salah satu pemecahan masalah dalam mengkategorikan dokumen teks dapat diselesaikan dengan menggunakan metode *text mining* yaitu klasifikasi.

Penelitian ini menggunakan algoritma *Cosine Similarity* untuk melakukan klasifikasi sesuai kemiripan teks *abstract* jurnal. *Cosine Similarity* telah banyak digunakan untuk melakukan pengklasifikasian teks seperti pengklasifikasian *tweet* populer (Ahmed, Razzaq, & Qamar, 2013), pengklasifikasian pertanyaan ujian (Jayakodi, Bandara, & Meedeniya, 2016), pengklasifikasian jawaban ujian (Saipech & Seresangtakul, 2018), pengklasifikasian komentar mahasiswa pada sistem evaluasi pembelajaran (Muhammad Habibi & Sumarsono, 2018) serta untuk pengklasifikasian dokumen *text* (Kadhim, Cheah, Ahamed, & Salman, 2014).

Selain algoritma *Cosine Similarity*, Penelitian ini juga menggunakan Algoritma *Support Vector Machine* (SVM) sebagai pembanding. SVM dikenal sebagai metode yang memiliki nilai akurasi yang sangat baik untuk pengklasifikasian data teks. Salah satu penerapan SVM untuk klasifikasi teks yaitu, klasifikasi judul skripsi (Hidayatullah & Maarif, 2016), dan klasifikasi komentar mahasiswa (Muhammad Habibi, 2017).

Tujuan penelitian ini adalah membuat sebuah model klasifikasi yang dapat mengklasifikasikan jurnal secara otomatis menggunakan algoritma *Cosine Similarity* dan *Support Vector Machine* dalam proses pengklasifikasiannya dan menggunakan metode pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode pembobotan menggunakan TF-IDF sudah banyak digunakan dalam pemrosesan data teks, seperti yang digunakan untuk pengolahan data *hashtag* pada *caption* Instagram (Muhamad Habibi & Cahyo, 2019) serta untuk analisis konten jejaring sosial twitter (Muhammad Habibi, 2018). Objek penelitian ini adalah *abstract* pada jurnal ilmiah. Jurnal akan diklasifikasikan sesuai dengan rumpun ilmu bidang keahlian dari *reviewer*. Sehingga diharapkan model klasifikasi yang dihasilkan pada penelitian ini dapat membantu meringankan kegiatan plotting *reviewer* pada editorial jurnal.

2. METODE PENELITIAN

2.1 DATASET

Data *abstract* jurnal yang digunakan dalam penelitian ini terdiri dari empat kelas bidang ilmu yaitu Sistem Cerdas, *Data Mining*, *Image Processing* dan Jaringan. Sebanyak 210 data jurnal yang digunakan dalam penelitian ini. Adapun detail *dataset* yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Jumlah Pengguna Internet di Indonesia

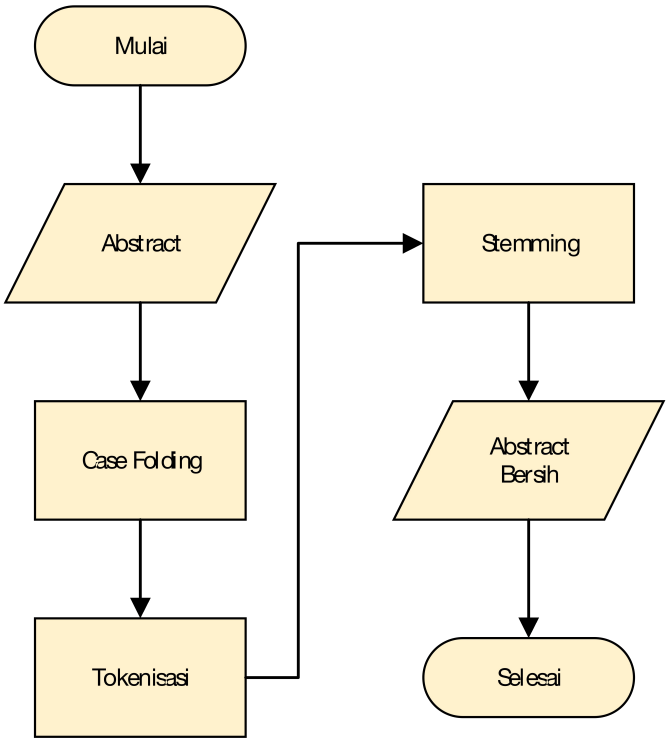
No	Kelas Bidang Ilmu	Jumlah Data
1	*Data Mining*	37
2	Sistem Cerdas	82
3	*Image Processing*	36
4	Jaringan	55
Total		210

2.2 PREPROCESSING

Penelitian ini terdiri dari beberapa tahapan, tahapan awal dalam penelitian ini adalah *preprocessing*. *preprocessing* data merupakan proses mempersiapkan dan membersihkan data teks sebelum teks dilakukan analisis (Haddi, Liu, & Shi, 2013). Adapun *flowchart preprocessing* dapat dilihat pada Gambar 1.

Tahapan *preprocessing* pada penelitian ini memiliki perbedaan dengan tahapan *preprocessing* yang dilakukan pada data teks media sosial. Data yang digunakan dalam penelitian ini merupakan data *abstract* bahasa Inggris yang terdapat pada jurnal. Berbeda dengan data teks sosial media, data teks *abtract* pada jurnal memiliki karakteristik kalimat yang baku dan sesuai dengan kaidah bahasa yang benar sehingga tidak terlalu banyak mengandung kata-kata tidak baku. Adapun tahapan *preprocessing* yang akan dilakukan dalam penelitian ini diantaranya adalah:

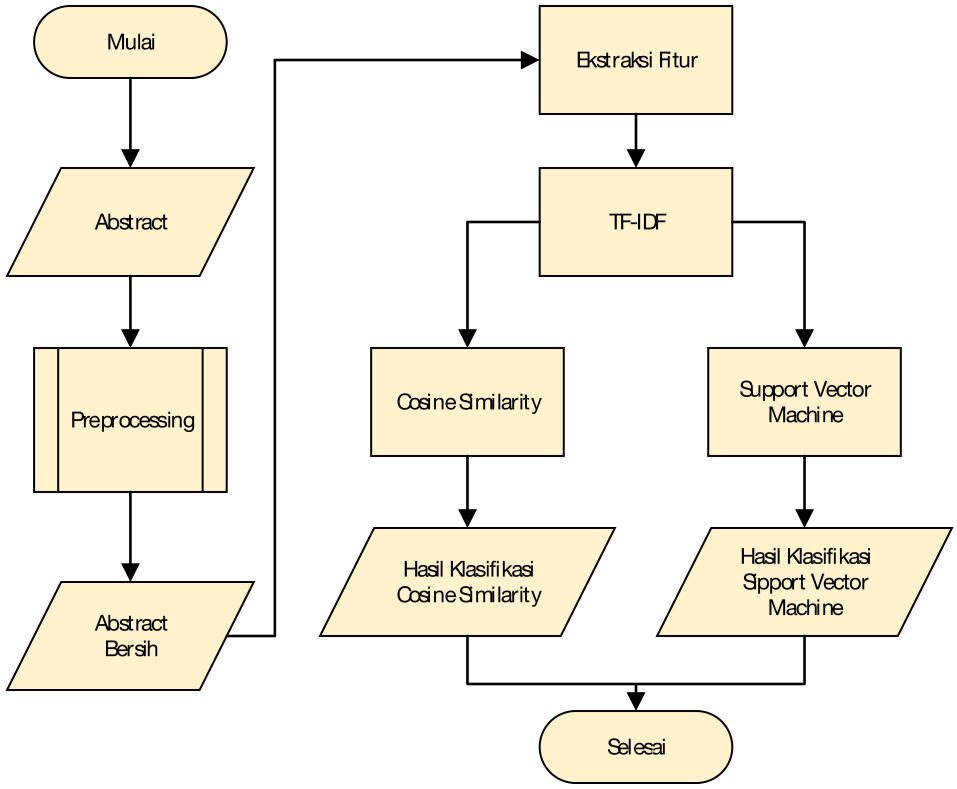
a. *Case Folding* yaitu proses untuk mengubah huruf kecil pada teks *abstract*.
b. Tokenisasi yaitu proses untuk membagi teks *abstract* ke dalam token.
c. *Stemming* yaitu proses mengubah kata berimbuhan menjadi kata dasar.



Gambar 1*. Flowchart Preprocessing*.

2.3 KLASIFIKASI

Tahapan selanjutnya setelah proses *preprocessing* adalah ekstraksi fitur, ekstraksi fitur bertujuan untuk mengidentifikasi entitas yang dirujuk (Siqueira & Barros, 2010). Pada penelitian ini, fitur yang digunakan adalah *Term Frequency – Inverse Document Frequency* (TF-IDF). TF-IDF terdiri dari dua buah nilai komponen yaitu *term-frequency* dan *inverse document frequency*. Skema pembobotan TF-IDF memberikan bobot *term* dalam suatu dokumen (Manning, Raghavan, & Schutze, 2009). Setelah didapatkan nilai TF-IDF, langkah selanjutnya adalah proses klasifikasi menggunakan *Cosine Similarity* dan *Support Vector Machine*. Flowchart proses klasifikasi dapat dilihat pada Gambar 2.



Gambar 2. Flowchart proses klasifikasi.

2.4 METODE EVALUASI

Estimasi tingkat kesalahan prediksi diperlukan untuk mengevaluasi kinerja model klasifikasi yang sudah dibuat. *Cross validation* dapat digunakan untuk memperkirakan kesalahan prediksi (Fushiki, 2011). Dalam pendekatan *cross validation*, setiap *record* digunakan beberapa kali dalam jumlah yang sama untuk pelatihan dan untuk pengujian.

Metode *k-fold cross-validation* mensegmentasi data ke dalam *k* partisi berukuran sama. Pada metode ini salah satu dari partisi dipilih untuk pengujian, sedangkan sisanya digunakan untuk pelatihan. Prosedur ini diulang sebanyak *k* kali sehingga setiap partisi digunakan untuk pengujian tepat satu kali. Total *error* ditentukan dengan menjumlahkan *error* untuk semua *k* proses tersebut. (Muhammad Habibi, 2017).

Perhitungan validasi hasil klasifikasi dapat diukur menggunakan *precision, recall*, dan *harmonic mean* dari *precision* dan *recall* yakni *F-score* (Dermawan, 2016). Pengujian dengan *precission* dan *recall* pada suatu entitas menunjukan hasil yang baik sehingga dapat meningkatkan nilai *F-score* (Cahyo, 2017). *Precision* merupakan persentase model klasifikasi dapat melakukan pelabelan benar dari label yang dikenali. *Recall* merupakan persentase seberapa banyak label dapat dikenali oleh model klasifikasi. Sedangkan *F-score* merupakan penghitungan evaluasi temu kembali informasi yang mengkombinasikan *recall* dan *precision*.

3. HASIL DAN PEMBAHASAN

3.1 HASIL AKURASI MODEL KLASIFIKASI

Proses perhitungan akurasi dilakukan dalam 10 kali percobaan menggunakan *K-fold cross validation*. Hasil perhitungan akurasi dapat dilihat pada Tabel 2.

Tabel 2. Hasil Perhitungan Akurasi Model Klasifikasi

Percobaan	Akurasi	
	Cosine Similarity	***Support Vector Machine***
1	0,81	0,71
2	0,62	0,76
3	0,67	0,57
4	0,67	0,81
5	0,57	0,67
6	0,67	0,76
7	0,57	0,67
8	0,62	0,76
9	0,48	0,81
10	0,48	0,95
Rata-rata	**0,61**	**0,75**

Berdasarkan Tabel 2, Hasil percobaan menunjukkan bahwa model klasifikasi yang dibangun dengan menggunakan algoritma *Cosine Similarity* memiliki nilai akurasi rata-rata dalam 10 kali percobaan yaitu 61%. Sementara itu, medel klasifikasi yang dibangun menggunakan algoritma *Support Vector Machine* memiliki nilai akurasi rata-rata dalam 10 kali percobaan yaitu 75%. Pada penelitian ini didapatkan bahwa akurasi algoritma *Support Vector Machine* memiliki akurasi yang lebih baik dibandingkan dengan algoritma *Cosine Similarity*.

3.2 PERHITUNGAN *PRECISION*, *RECALL* DAN *F-SCORE*

Hasil perhitungan evaluasi *precision, recall* dan *f-score* menggunakan algoritma *Cosine Similarity* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Perhitungan *Precision, Recall* dan *F-score Cosine Similarity*

Percobaan	***Precision***	***Recall***	***F-score***
1	0,86	0,88	0,85
2	0,73	0,67	0,61
3	0,68	0,74	0,62
4	0,81	0,62	0,67
5	0,68	0,54	0,50
6	0,64	0,73	0,66
7	0,63	0,63	0,56
8	0,53	0,52	0,51
9	0,44	0,43	0,43
10	0,47	0,46	0,43
Rata-rata	**0,65**	**0,62**	**0,58**

Berdasarkan hasil yang ditunjukkan pada Tabel 3, diketahui bahwa algoritma *Cosine Similarity* menghasilkan nilai rata-rata *precision* sebesar 65%, *recall* sebesar 63%, sedangkan *f-score*

yang dihasilkan adalah 58%. Hasil perhitungan evaluasi menggunakan algoritma *Support Vector Machine* dapat dilihat pada Tabel 4. Pada tabel tersebut, dapat diketahui bahwa hasil nilai rata-rata *precision* sebesar 72%, *recall* sebesar 69%, sedangkan *f-score* sebesar 67%.

Tabel 4. Hasil Perhitungan *Precision, Recall* dan *F-score* SVM

Percobaan	*Precision*	*Recall*	*F-score*
1	0,72	0,71	0,71
2	0,56	0,58	0,56
3	0,59	0,51	0,47
4	0,89	0,78	0,81
5	0,37	0,45	0,40
6	0,88	0,75	0,76
7	0,67	0,67	0,64
8	0,78	0,69	0,70
9	0,86	0,78	0,75
10	0,92	0,97	0,94
Rata-rata	**0,72**	**0,69**	**0,67**

Hasil perhitungan *precision*, *recall* dan *f-score* dari algoritma *Cosine Similarity* dan *Support Vector Machine* menunjukkan bahwa hasil pengujian model klasifikasi yang dibangun menggunakan *Support Vector Machine* memiliki hasil *precision*, *recall* dan *f-score* lebih baik dibandingkan dengan *Cosine Similarity*.

3.3 PERHITUNGAN *PRECISION, RECALL* DAN *F-SCORE* UNTUK TIAP LABEL

Hasil perhitungan *precision*, *recall* dan *f-score* untuk masing-masing *class label* menggunakan *Cosine Similarity* dan *Support Vector Machine* secara berturut-turut dapat dilihat pada Tabel 5 dan Tabel 6.

Tabel 5. Hasil *Precision, Recall* dan *F-score* tiap label menggunakan *Cosine Similarity*

Class Label	*Precision*	*Recall*	*F-score*
Sistem Cerdas	0,60	0,76	0,65
Data Mining	0,71	0,53	0,54
Image Processing	0,50	0,51	0,45
Jaringan	0,78	0,68	0,70

Tabel 6. Hasil *Precision*, *Recall* dan *F-score* tiap label menggunakan SVM

Class Label	*Precision*	*Recall*	*F-score*
Sistem Cerdas	0,67	0,89	0,75
Data Mining	0,64	0,46	0,50
Image Processing	0,69	0,63	0,65
Jaringan	0,89	0,77	0,81

Berdasarkan hasil Tabel 5, hasil *Precision, Recall* dan *F-score* untuk setiap label menggunakan *Cosine Similarity* didapatkan bahwa kategori *Image Processing* memiliki nilai *precision*, *recall* dan *f-score* paling rendah dibandingkan dengan kategori yang lain. Sedangkan hasil *Precision*, *Recall* dan *F-score* untuk setiap label menggunakan *Support Vector Machine* didapatkan bahwa kategori *Data Mining* memiliki nilai *precision, recall* dan *f-score* paling rendah dibandingkan dengan kategori yang lain.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, maka diperoleh kesimpulan bahwa, penelitian ini berhasil membuat model yang dapat mengklasifikasikan jurnal secara otomatis menggunakan algoritma *Cosine Similarity* dan *Support Vector Machine* dan menggunakan metode pembobotan TF-IDF. Hasil akurasi pengujian metode *Cosine Similarity* diperoleh sebesar 61% sedangkan metode *Support Vector Machine* didapatkan akurasi sebesar 75%. Metode *Support Vector Machine* menghasilkan akurasi performansi yang lebih baik dari pada metode *Cosine Similarity*. Hasil perhitungan nilai *precision*, *recall*, dan *f-score* diketahui bahwa metode *Support Vector Machine* menghasilkan nilai yang lebih baik, sejalan dengan nilai akurasi.

.

DAFTAR PUSTAKA

Ahmed, H., Razzaq, M. A., & Qamar, A. M. (2013). Prediction of popular tweets using Similarity Learning. *ICET 2013 - 2013 IEEE 9th International Conference on Emerging Technologies*. https://doi.org/10.1109/ICET.2013.6743524

Cahyo, P. W. (2017). *Model Monitoring Sebaran Penyakit Demam Berdarah di Indonesia Berdasarkan Analisis Pesan Twitter*. Universitas Gadjah Mada Yogyakarta.

Dermawan, R. (2016). *Klasifikasi Tweet dan Pengenalan Entitas Bernama pada Tweet Bencana Dengan Support Vector Machine*. Universitas Gadjah Mada.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Habibi, Muhamad, & Cahyo, P. W. (2019). Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(4), 399–408.

Habibi, Muhammad. (2017). *Analisis Sentimen dan Klasifikasi Komentar Mahasiswa pada Sistem Evaluasi Pembelajaran Menggunakan Kombinasi KNN Berbasis Cosine Similarity dan Supervised Model*. *Departemen Ilmu Komputer dan Elektronika, Fakultas Matematika dan Ilmu Pengetahuan Alam*. Universitas Gadjah Mada.

Habibi, Muhammad. (2018). Analisis Konten Jejaring Sosial Twitter dalam Kasus Pemilihan Gubernur DKI 2017. *Teknomatika*, *11*(1), 31–40.

Habibi, Muhammad, & Sumarsono. (2018). Implementation of Cosine Similarity in an automatic classifier for comments. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *3*(2), 38–46.

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, *17*, 26–32. https://doi.org/10.1016/j.procs.2013.05.005

Hidayatullah, A. F., & Maarif, M. R. (2016). Penerapan Text Mining dalam Klasifikasi Judul Skripsi. In *Seminar Nasional Aplikasi Teknologi Informasi (SNATi) Agustus* (pp. 1907–5022). Yogyakarta.

Jayakodi, K., Bandara, M., & Meedeniya, D. (2016). An automatic classifier for exam questions with WordNet and Cosine similarity. *2nd International Moratuwa Engineering Research Conference, MERCon 2016*, 12–17. https://doi.org/10.1109/MERCon.2016.7480108

Kadhim, A. I., Cheah, Y. N., Ahamed, N. H., & Salman, L. A. (2014). Feature extraction for co-occurrence-based cosine similarity score of text documents. *2014 IEEE Student Conference on Research and Development, SCOReD 2014*, 2–5.

https://doi.org/10.1109/SCORED.2014.7072954

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press. https://doi.org/10.1109/LPT.2009.2020494

Saipech, P., & Seresangtakul, P. (2018). Automatic Thai Subjective Examination using Cosine Similarity. *ICAICTA 2018 - 5th International Conference on Advanced Informatics: Concepts Theory and Applications*, 214–218. https://doi.org/10.1109/ICAICTA.2018.8541276

Siqueira, H., & Barros, F. (2010). A Feature Extraction Process for Sentiment Analysis of Opinions on Services. *Proceedings of the III International Workshop on Web and Text Intelligence (WTI)*.