Klasifikasi Berita Menggunakan Vector Space Model News Classification Using Vector Space Model

Aulia Tegar Rahman[1], Astika Wulansari[2]
Program Studi Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta, Jl. Padjajaran, Ring Road Utara, Kel. Condongcatur, Kec. Depok, Kab. Sleman, Daerah Istimewa Yogyakarta, Indonesia

Email: aulia.tegar.rahman@students.amikom.ac.id, astikaws@students.amikom.ac.id

ABSTRAK

Portal berita diluncurkan dengan tujuan memenuhi kebutuhan hak masyarakat akan informasi yang akurat, lengkap, bermanfaat dan terkini. Detik.com salah satu portal web yang menyampaikan berita dan artikel secara daring di Indonesia. Detik.com menampilkan *breaking news*, sehingga *vivid description* ini menjadikan detik.com berkembang pesat menjadi situs informasi digital yang populer. Artikel berita tersusun dari kata-kata terpilih yang mewakili sebuah berita. Kata yang sering muncul dalam berita tersebut menjadi kata kunci sehingga temu kembali informasi mampu melakukan klasifikasi berita. Proses penambahan *crawling data* dan proses pelatihan dapat meningkatkan kecepatan dan ketepatan klasifikasi berita pada portal detik.com. Data yang saat ini diproses memberikan evaluasi untuk meningkatkan akurasi. Data pelatihan tentang klasifikasi berita menggunakan *Vector Space Model (VSM)* mendapatkan akurasi yang lebih bagus yaitu 40%.

Kata kunci: *berita; klasifikasi; vector space model (VSM).*

ABSTRACT

The news portal was launched with the aim of fulfilling the public's right to obtain accurate, complete, useful and up-to-date information. Detik.com is a web portal containing online news and articles in Indonesia. Detik.com makes breaking news, relying on this vivid description, makes detik.com the most popular digital information site. News articles are composed of selected words that represent a story. Words that often appear in the news become keywords so that information retrieval is able to classify news. The process of adding data crawling and training processes can increase the speed and accuracy of news classification on the detik.com portal. The currently processed data provides evaluations to improve accuracy. Training data on news classification using the Vector Space Model (VSM) gets a better accuracy of 40%

Keywords: *classification; news; vector space model (VSM).*

1. PENDAHULUAN

Klasifikasi Berita Menggunakan *Vector Space Model* (VSM)

Keunggulan layanan informasi berita dan artikel berbasis web adalah memberikan kemudahan dan kenyamanan dalam mengakses berita terkini dalam berbagai kategori berita setiap hari sesuai kebutuhan masyarakat. Portal berita diluncurkan dengan tujuan memenuhi kebutuhan hak masyarakat akan informasi yang akurat, lengkap, bermanfaat dan terkini. Karena tuntutan pemenuhan kebutuhan masyarakat akan kecepatan berita, tak jarang melakukan kekeliruan dalam penulisan sehingga berpengaruh dalam kualitas dan index berita. Index berita merupakan kumpulan berita terkini berdasarkan kategori berita, dan tanggal publikasi (Londo, Greeley et al ,2019).

Di Indonesia, detik.com termasuk salah satu portal web yang menyampaikan berita dan artikel secara daring. Detik.com menampilkan *breaking news,* sehingga *vivid description* ini menjadikan detik.com melesat sebagai situs informasi digital yang populer. Pada awal kemunculannya detik.com setiap hari menerima 30.000 hits hingga saat ini mencapai lebih dari 2.500.000 hits setiap hari. Detik.com menggunakan alat ukur *page view* untuk mengetahui ukuran seberapa besar potensi yang dimiliki sebuah website. *Page view* adalah jumlah halaman yang diakses setiap hari. Saat ini detik.com menempati posisi empat tertinggi jumlah 3.000.000 page view setiap hari. Dalam memproduksi berita yang cepat dan akurat, tak jarang jurnalis menulis berita langsung ketika peristiwa sedang berlangsung sekaligus

melakukan perekaman sesuai dengan instruksi redaksi. Apalagi jika isu yang dibicarakan penting dan harus diterbitkan segera mungkin. Artikel yang akan diterbitkan sebelumnya akan disusun dan diperiksa oleh editor untuk dilakukan penyuntingan.
Temu kembali informasi mampu menentukan secara cepat dan akurat dalam menyusun dokumen yang telah didapat ditampilkan berurut dari dokumen yang relevansi tinggi ke relevansi rendah. Artikel berita tersusun dari kata-kata terpilih yang mewakili sebuah berita. Kata yang sering muncul dalam berita tersebut menjadi kata kunci sehingga temu kembali informasi mampu melakukan klasifikasi berita berdasarkan kata yang sering muncul. Klasifikasi berita ini penting dalam penyajian berita berdasarkan indexs. Kecepatan dalam penyajian berita menuntut ketepatan dan keakuratan dalam klasifikasi berita.
Berdasarkan latar belakang tersebut, penulis melakukan penelitian "Klasifikasi Berita Menggunakan *Vector Space Model* (SVM)".

2. TINJUAN PUSTAKA

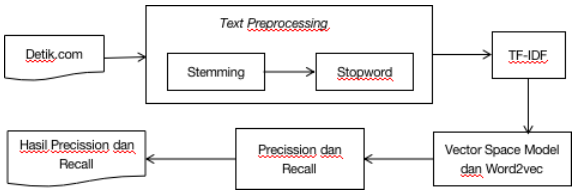
Pada penelitian sebelumnya Suprianto dan Muhammad Fadlan menggunakan klasifikasi *Vector Space Model* dan *Naive Bayes* dalam mengevaluasi kinerja dosen dari mahasiswa (Suprianto et al., 2020).

Penelitian yang lain Hay Man Oo dan Win Pa Pa menggunakan *Vector Space Model* dalam temu kembali informasi dimana dokumen mewakili sebagai vektor dalam ruang n-dimensi dimana setiap dimensi mewakili *term* dan diukur melalui sudut consinus antara 2 vektor (Oo et al., 2020).

Sementara itu Omid Shahmirzadi, Adam Lugowski, dan Kenneth Younge melakukan evaluasi berbagai *Vector Space Model* dengan TF-IDF pada pengukuran otomatis kesamaan teks simantik (Shahmirzadi et al., 2019).

3. METODOLOGI PENELITIAN

Metodologi penelitian berisi bahan utama yang digunakan dalam penelitian dan metode yang digunakan untuk memecahkan permasalahan termasuk metode analisis dan penjelasan gambarnya.
Tahapan yang dilakukan pada penelitian ini sehingga mendapatkan hasil klasifikasi yang optimal meliputi Pengumpulan data, Text Preprocessing, Vektorisasi, Klasifikasi dengan *metode Naive Bayes dan Space Vector Model (SVM)* sesuai pada Gambar 1.



Gambar 1 Flowchart Metodologi

Berdasarkan Gambar 1. *Flowchart* Metodologi dijelaskan sebagai berikut *:*

Metode Pengumpulan Data
Dalam melakukan penelitian ini, pengumpula n data dilakukan menggunakan web crawling untuk mengindeks informasi pada portal berita detik.com dengan menggunakan newspaper3k *python library*. Dengan ukuran dataset 201, menggunakan 100 data train dan 33 data uji.

Text Preprocessing
Penelitian ini menggunakan teknik *text preprocessing*, dengan menghilangkan tanda baca dan angka terlebih dahulu. Proses *Text Preprocessing* ini untuk memilih kata sebagai indeks. Indeks merupakan kata-kata yang mewakili sebuah dokumen dan digunakan untuk membuat pemodelan temu kembali informasi. *Text preprocessing* yang digunakan *stemming* dan *stopword*. Stemming merupakan proses mencari kata dasar. Menghilangkan imbuhan (awalan, sisipan, dan akhiran) kemudian menggantikan bentuk kata tersebut menjadi kata yang sesuai Bahasa Indonesia yang baik dan benar. *Stopword* adalah kata umum yang biasanya muncul dalam jumlah yang besar dan dianggap tidak memiliki makna. Berguna untuk mengurangi jumlah kata yang diproses.

Vektorisasi
Proses vektorisasi menggunakan *Vector Space Model (VSM)* pada penelitian ini dimana melakukan pendekatan natural yang berbasis vektor pada setiap kata sebuah dimensi spasial.
Pembobotan
Pembobotan menggunakan TF-IDF *(Term Frequently-Inverse Document Frequency) u*ntuk mengetahui kata mana yang paling penting. TF merupakan frekuensi kemunculan *term* i pada dokumen j dibagi dengan total *term* pada dokumen j. Ditulis dalam bentuk :

$$tf_{ij} = \frac{f_d(i)}{\max_{j \in d} f_d(j)} \quad (1)$$

Inverse document frequency (IDF) merupakan proses pengurangan bobot suatu *term* jika frekuensi muncul tersebar di seluruh dokumen, dituliskan dalam bentuk:

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (2)$$

Naive Bayes (NB)

Metode klasifikasi yang digunakan untuk klasifikasi teks berdasarkan kemungkinan. Metode NB melibatkan 2 tahap yaitu (Singh, Gurinder., et al 2019)*:*

Tahap pelatihan. Melakukan proses analisis terhadap contoh dokumen terhadap pemilihan kata, dan kemungkinan kata yang muncul pada contoh dokumen akan menjadi representasi dokumen tersebut.

Tahap klasifikasi. Nilai sebuah dokumen berdasarkan kata yang muncul.

***Vector Space Model* (VSM)**

Pendekatan natural berbasis vektor dari setiap kata dalam suatu dimensi spasial. Pada VSM sebuah kata direpresentasikan dengan sebuah dimensi ruang vektor. (Samant, S., S., et al, 2019; Shahmirzadi, O., 2019; Parida, U, 2019; Shahmirzadi, O., 2019; Oo, H., M., et al, 2020)

Dalam VSM, kumpulan dokumen direpresentasikan menjadi sebuah matriks *term-document* (atau matrik *term frequency*). Tiap sel matrik sesuai dengan bobot yang diberikan suatu *term* pada dokumen yang dipilih. Nilai nol dapat diartikan bahwa *term* tidak ada dalam dokumen.

Tabel 1. Matriks Term document

	T_1	T_2	T_3	$T_{...}$	T_t
D_1	W_{11}	W_{21}	W_{31}	$\ldots_{..}$	T_{t1}
D_2	W_{12}	W_{22}	W_{32}	$\ldots_{..}$	T_{t2}
D_3	W_{13}	W_{23}	W_{33}	$\ldots_{..}$	T_{t3}
$D_{...}$	$\ldots_{...}$	$\ldots_{..}$	$\ldots_{..}$	$\ldots_{..}$	$T_{t..}$
D_n	W_{1n}	W_{2n}	W_{3n}	$\ldots_{..}$	T_{tn}

Precision dan Recall

Temu kembali informasi merupakan pencarian berdasarkan kata kunci dan tingkat kesamaan pada dokumen yang tidak terstruktur. Untuk mengukur kualitas penemuan teks menggunakan :

Precision, jumlah dokumen yang ditemukan dan dianggap relevan untuk temu kembali informasi dari jumlah keseluruhan dokumen yang relevan.

Precision=

$$\frac{jumlah\ dokumen\ relevan\ yang\ terpanggil\ (a)}{Jumlah\ dokumen\ yang\ terpanggil\ dalam\ pencarian\ (a+b)} \text{ x } 100$$

Recall, kemampuan sistem memanggil kembali dokumen yang dianggap relevan dari pangkalan data (database).

Recall=

$$\frac{Jumlah\ dokumen\ relevan\ yang\ terpanggil\ (a)}{Jumlah\ dokumen\ relevan\ yang\ ada\ di\ database\ (a+c)} \text{ x } 100$$

Tabel 2. Matriks precision dan recall

	Relevan	Tidak relevan	Total
Temu Kembali	a *(hits)*	b *(noise)*	(a+b)
Bukan Temu Kembali	c *(noise)*	d *(reject)*	(c+d)
Total	(a+c)	(b+d)	(a+b+c+d)

4. PEMBAHASAN

Terlebih dahulu melakukan *import library sklearn* dan library sastrawi. *Library sklearn* digunakan untuk mengelompokkan satu set objek yang tidak berlabel, perkiraan hubungan antar variabel, dan menentukan klasifikasi pengamatan baru. Sedangkan library sastrawi memungkinkan untuk mengurangi kata-kata yang terinfeksi dalam bahasa indonesia ke bentuk dasarnya. Kumpulan data hasil *web crawling* berupa data tidak terstruktur terlebih dahulu dilakukan konversi menjadi data terstruktur sesuai dengan kebutuhan klasifikasi berita. Proses ini menghapus tanda baca dan angka. Selanjutnya pembagian data pelatihan dan data pengujian. Ukuran data pelatihan 100 dan data pengujian 33. Kemudian mengumpulkan data yang sebagian besar memiliki elemen yang tidak membawa informasi apapun *(sparse data).* Ukuran *sparse data* 133 x 2. Tahapan *text preprocessing* ini mengubah kata ke bentuk dasar sekaligus sehingga dapat mengurangi jumlah kata tersebut. Tahapan

tersebut *stemming* dan *stopword.* Teknik *stemming* digunakan untuk meningkatkan performa dengan cara menemukan variasi token. Keuntungan *stemming* adalah efisiensi dan kompresi file. *Stopword* merupakan proses menghilangkan kata yang sering muncul tetapi tidak relevan terhadap dokumen.

Naive Bayes (NB).

Terdapat kategori berita, yaitu news detik, news detik indeks, hoaxornot detik, Inet detik, foto detik, hot detik, pasangmata detik, wolipop detik, 20 detik, health detik, food detik, travel detik, finance detik, oto detik, event detik, sport detik dengan jumlah label 0 - 14, data pelatihan 106. Sedangkan data pengujian 27. Setiap data pelatihan dan data pengujian melewati *text preprocessing*, vektorisasi, pembobotan dan *precision recall*. Hasil pembobotan NB.

```
[[1 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 1 0 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 1 0 1 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 0 0 0 1 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 2 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 1 0 0 0 0]
 [1 0 1 0 0 1 1 0 1 3 0 1 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [1 1 0 0 0 2 1 0 0 0 2 0 1]]
```

Gambar 2 Matriks pembobotan NB.

Hasil penghitungan *precision* dan *recall* NB seperti gambar 3 berikut.

	precision	recall	f1-score	support
0	0.33	1.00	0.50	1
1	0.00	0.00	0.00	2
2	0.00	0.00	0.00	0
3	1.00	0.50	0.67	2
4	0.00	0.00	0.00	2
5	0.00	0.00	0.00	0
6	0.50	1.00	0.67	2
7	0.00	0.00	0.00	0
8	0.50	1.00	0.67	1
9	1.00	0.33	0.50	9
10	0.00	0.00	0.00	0
11	0.00	0.00	0.00	0
12	1.00	0.12	0.22	8
accuracy			0.33	27
macro avg	0.33	0.30	0.25	27
weighted avg	0.77	0.33	0.37	27

```
Akurasi Naive Bayes -> 33.33333333333333
```

Gambar 3 Nilai precision dan recall NB..

Vector Space Model (VSM).

Data pelatihan dan data pengujian serupa yaitu data pelatihan 106 dan data pengujian 27. Sedangkan kata kunci yang diambil sembarang. Analisis VSM bertumpu pada nilai *precision* dan *recall* yang diperoleh. Hasil pembobotan VSM seperti pada gambar 4 berikut.

```
[[1 0 0 0 0 0 0 0 0 0 1 1 0]
 [0 0 1 0 1 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 1 0 1 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 0 0 0 1 0]
 [0 0 0 0 0 2 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 2 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 1 0 0 0 0]
 [1 0 1 0 0 0 1 0 1 3 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 0 0]
 [1 1 0 0 0 1 1 0 0 0 1 0 1]]
```

Gambar 4 Matriks pembobotan VSM

Hasil penghitungan *precision* dan *recall* VSM seperti gambar 5 berikut.

	precision	recall	f1-score	support
0	0.33	0.33	0.33	3
1	0.00	0.00	0.00	2
2	0.00	0.00	0.00	0
3	1.00	0.50	0.67	2
4	0.00	0.00	0.00	2
5	0.67	1.00	0.80	2
6	0.50	1.00	0.67	2
7	0.00	0.00	0.00	0
8	0.50	1.00	0.67	1
9	1.00	0.43	0.60	7
10	0.00	0.00	0.00	0
11	0.00	0.00	0.00	0
12	1.00	0.17	0.29	6
accuracy			0.41	27
macro avg	0.38	0.34	0.31	27
weighted avg	0.70	0.41	0.44	27

```
Akurasi SVM ->  40.74074074074074
```

Gambar 4 Matriks pembobotan VSM

4. KESIMPULAN

Proses penambahan crawling data dari portal berita detik.com dan proses pelatihan dapat meningkatkan kecepatan dan ketepatan klasifikasi berita pada portal detik.com. Data yang saat ini diproses memberikan evaluasi untuk meningkatkan akurasi. Data pelatihan tentang klasifikasi berita menggunakan *Vector Space Model (VSM)* mendapatkan akurasi yang lebih bagus yaitu 40%.

DAFTAR PUSTAKA

Bhavadharani, M., et al(2019), 'Performance Analysis Of Ranking Models In Information Retrieval' in Proceedings of the Third International Conference on Trends in Electronics and Informatics.

doi: 10.1109/ICOEI.2019.8862785

Martino, A., et al (2020), 'An Ecology-based Index for Text Embedding and Classification' in International Joint Conference on Neural Networks (IJCNN). doi: 10.1109/IJCNN48605.2020.9207299

Oo, H., M., et al (2020), 'Myanmar News Retrieval in Vector Space Model using Cosine Similarity Measure' in IEEE Conference on Computer Applications(ICCA). doi : 10.1109/ICCA49400.2020.9022845

Parida, U., (2019), 'Ranking of Odia Text Document relevant to User Query using Vector Space Model' in International Conference on Applied Machine Learning (ICAML). Doi : 10.1109/ICAML48257.2019.00039

Miao,Fang., et al(2019), 'News Text Classification Based on Machine Learning Algorithm' in International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). DOI:10.1109/IHMSC.2018.10117

Londo, Greeley , et al (2019), 'A Study of Text Classification for Indonesian News Article' in International Conference of Artificial Intelligence and Information Technology (ICAIIT). DOI:10.1109/ICAIIT.2019.8834611

Singh, Gurinder., et al (2019), 'Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification' in International Conference on Automation, Computational and Technology Management (ICACTM). DOI:10.1109/ICACTM.2019.8776800

Samant, S., S., et al (2019), 'Improving Term Weighting Schemes for Short Text Classification in Vector Space Model' in IEEE Access (Volume: 7) doi : 10.1109/ACCESS.2019.2953918

Shahmirzadi, O., (2019), 'Text Similarity in Vector Space Models: A Comparative Study' in 18th IEEE International Conference on Machine Learning and Applications (ICMLA). doi : 10.1109/ICMLA.2019.00120.

Suprianto, et al (2020), 'Retrieval Information Using Generalized Vector Space Models And Sentiment Analysis Using Naïve Bayes Classifier For Evaluation Of Lecturers By Students' in Fifth International Conference on Informatics and Computing (ICIC). Doi : 10.1109/ICIC50835.2020.9288584.

Wahyudi, E., (2019), 'Information Retrieval System for Searching JSON Files with Vector Space Model Method' in international Conference of Artificial Intelligence and Information Technology (ICAIIT). doi : 10.1109/ICAIIT.2019.8834457.