Gunung Djati Conference Series, Volume 3 (2021)
Mini Seminar Kelas *Data Mining* 2020
ISSN: 2774-6585
Website: https://conferences.uinsgd.ac.id/gdcs

Klasifikasi Kategori Berita menggunakan Algoritma *Support Vector Machine*

(*News Category Classification using Support Vector Machine Algorithm*)

Nisa Eka Juliana[1], Faridah Dewi Khansa[2], Aaz M Hafidz Azis[3], Rafli Indra Gunawan[4], Nurul Dwi Cahya[5]

[1]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050116@student.uinsgd.ac.id
[2]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050042@student.uinsgd.ac.id
[3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050001@student.uinsgd.ac.id
[4]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050093@student.uinsgd.ac.id
[5]Teknik Informatika, UIN Sunan Gunung Djati Bandung, nuruldwicahya925@gmail.com

Abstrak

Pada jaman sekarang sudah banyak yang menggunakan sistem berbasis web untuk menyampaikan informasi dan berita secara real time. Namun, dalam membagi berita ke dalam kategori-kategori tersebut masih ada yang dilakukkan secara manual, sehingga perlu memerlukan waktu yang cukup lama. Dari beberapan teknik yang ada teknik yang paling sering digunakan untuk klasifikasi konten berita adalah Support Vector Machine (SVM). Pada permasalahan yang kompleks atau permasalahan dengan parameter yang banyak, metode ini sangat baik untuk digunakan. Algoritma SVM melakukan klasifikasi secara supervised learning atau mempunyai input dan output yang telah dibentuk menjadi suatu model hubungan matematis yang dapat melakukan klasifikasi dan prediksi dari data yang sudah ada sebelumnya. Terdapat 2224 dataset dan 5 kategori dengan 70% data yang ditraining dan 30% data yang ditesting. Penelitian ini menghasilkan klasifikasi teks dalam bentuk kategori Teknologi, bisnis, sport, entertaimen, dan politik dari konten berita digital. Hasil klasifikasi diperoleh nilai akurasi mencapai 98.35 % dengan rata-rata presisi 90%, recall 98%, F1-Score 98% dan Support sebesar 668.

Kata kunci: data mining, klasifikasi berita, support vector machine

Abstract

Nowadays many have used web-based systems to convey information and news in real time. However, in dividing news into these categories, some are still done manually, so it takes a long time. Of the several existing techniques, the technique most often used for classification of news content is the Support Vector Machine (SVM). In complex problems or problems with many parameters, this method is very good to use. The SVM algorithm performs supervised learning classifications or has inputs and outputs that have been formed into a mathematical relationship model that can classify and predict existing data. There are 2224 datasets and 5 categories with 70% of the data being trained and 30% of the data being tested. This study produces text classifications in the form of technology, business, sports, entertainment, and political categories from digital news content. The classification results obtained an accuracy value of 98.35% with an average precision of 90%, a recall of 98%, an F1-score of 98% and a Support of 668.

Keywords: *data mining, news classification, support vector machine*

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

1 Pendahuluan

Berita yaitu suatu informasi yang didapatkan dari kejadian yang terjadi mengenai suatu hal, seiring berjalannya perkembangan jaman, berita bisa didapatkan dalam bentuk cetakan seperti koran, siaran di televisi, internet, bahkan bisa dari beberapa atau sekumpulan orang. Pada masa ini, berita dapat dilihat menggunakan internet seperti kompas.com, detik.com, kumparan, dan lain-lain yang merupakan salah satu website berita yang sering dikunjungi. Terdapat berbagai macam informasi yang bisa kita dapat dalam website tersebut. Terkadang, kita langsung menerima berita tanpa adanya penyeleksian informasi. Karena adanya hal seperti itu media informasi melakukan pengklasifikasian dengan mengkategorisasi terlebih dahulu sebelum dipublikasikan kepada khalayak umum. Pengklasifikasian tersebut bertujuan untuk memudahkan masyarakat untuk mencari informasi yang mereka inginkan.

SVM mampu mengidentifikasikan hyperplane terpisah yang dapat memaksimalkan margin antara dua kelas yang berbeda [1]. SVM memiliki kelebihan yaitu Algoritma supervised yang berupa klasifikasi dengan cara membagi data menjadi dua kelas menggunakan garis vektor yang disebut hyperplane. Sedangkan kekurangan SVM adalah memiliki kelemahan terhadap masalah pemilihan parameter atau fitur yang sesuai.

Klasifikasi atau pengkategorian teks adalah proses yan menempatkan sebuah dokumen ke suatu kategori atau kelasnya sesuai dengan karakteristik dari dokumen itu sendiri [2]. Di dalam text mining, klasifikasi menganalisis atau mempelajari himpunan dokumen teks preclasified agar mendapatkan suatu model atau fungsi dan hasilnya akan digunakan untuk pengelompokkan dokumen teks lain yang belum diketahui kelasnya dan di kelompokkan dalam satu atau lebih kelas pre-defined [3].

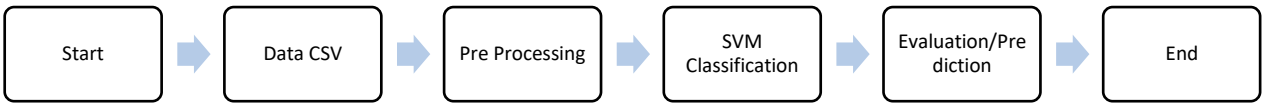
Support Vector Machines (SVM) secara signifikan mempengaruhi akurasi klasifikasi dikarenakan pengaturan parameter kernel dalam prosedur pelatihan SVM, bersama dengan pilihan fitur [4]. Pada algoritma Support Vector Machines ini terdapat metode yang berkaitan dengan metode pembelajaran, bertujuan untuk menyelesaikan sebuah permasalahah dari klasifikasi dan regresi [5], [6]. SVM secara konseptual bisa dikatakkan sebuah mesin linear yang dilengkapi dengan fitur khusus, dan berdasarkan minimisasi risiko struktural (SRM) metode dan teori belajar statistic [7], [8]. SVM ini adalah model linier yang menerapkan batasan kategori nonlinier dengan mengubah ruang contoh yang diberikan menjadi sebuah linear yang dipisahkan satu sampai pemetaan nonlinear [9].

Terdapat beberapa penelitian yang terkait dengan penelitian ini, antara lain: (1) hybrid algoritma Hidden Markov Model dan Support Vector Machine untuk klasifikasi berita dari website [10]; (2) klasifikasi berita berbahasa Indonesia dengan menggunakan SVM berbasis Particle Swam Optimization [11]; (3) kategorisasi dokumen teks berbahasa Arab dengan algoritma SVM [12]; dan analisis opini pada sosial media menggunakan algoritma SVM [13].

2 Metodologi

Di bagian ini, kami mendeskripsikan mengenai pengaplikasian atau implementasi algoritma Support Vector Machine (SVM) dalam mengklafisikasikan berita berdasarkan kategorinya [14] . Algoritma SVM melakukan klasifikasi secara *supervised learning* atau mempunyai input dan output yang telah dibentuk menjadi suatu model hubungan matematis yang dapat melakukan klasifikasi dan prediksi dari data yang sudah ada sebelumnya.

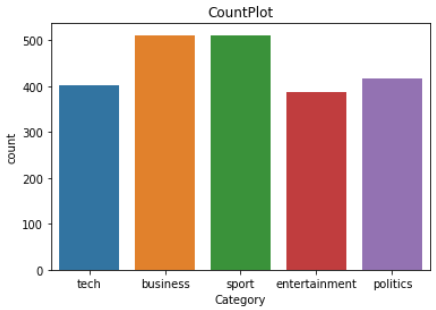
Adapun tahapan yang dijelaskan seperti berikut:



Gambar 1. Alur proses klasifikasi berita

Dalam proses pengumpulan data, dataset yang digunakan berasal dari kumpulan berita BBC News yang telah diberikan kategori atau label. Label ini terdiri atas Teknologi, bisnis, sport, entertaimen, dan politik. Yang kemudian dataset ini digunakan untuk training terhadap system agar dapat memprediksi berita-berita lainnya.

	category	text
0	tech	tv future in the hands of viewers with home th...
1	business	worldcom boss left books alone former worldc...
2	sport	tigers wary of farrell gamble leicester say ...
3	sport	yeading face newcastle in fa cup premiership s...
4	entertainment	ocean s twelve raids box office ocean s twelve...



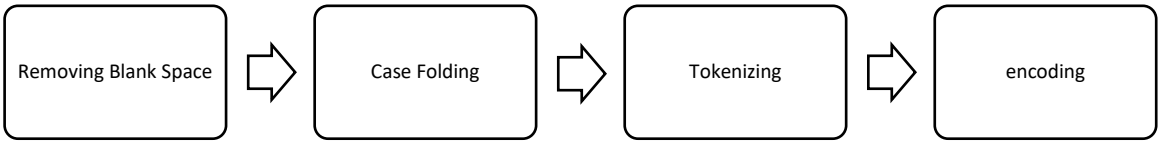
Gambar 2. Kumpulan Data berita dari BBC News

Tahapan preprocessing terhadap dataset ini merupakan proses pemilahan data menjadi data yang bisa langsung di lakukan training menggunakan algoritma svm, dalam melakukan proses ini terbagi menjadi 5 tahapan yaitu:

1. Removing blankspace.yaitu penghapusan spasi yang terdapat dalam kalimat sehingga menyisakan kata-kata yang diperlukan untuk data latih.
2. Casefolding, yaitu pengubahan penulisan menjadi lowcase pada semua kata/huruf.
3. Tokenizing, yaitu pemotongan pada setiap kalimat menjadi beberapa bagian. Setiap kalimat tersebut dikumpulkan dalam bentuk list.
4. Encoding, yaitu tahapan yang merubah kata-kata atau huruf menjadi bentuk vector atau numerik.

Pada tahapan klasifikasi menggunakan SVM ini, data dibagi menjadi dua bagian yaitu data train dan data test yang nantinya digunakan sebagai evaluasi dari kinerja SVM dalam melakukan proses klasifikasi. Data yang sudah dilatih yang digunakan pada tahapan ini sudah dilakukan melewati tahap processing yang sudah siap untuk dilakukan training.

Tahapan akhir pada proses ini yaitu evaluasi yang mengukur tingkat akurasi, f1-score dari proses klasifikasi yang terdapat pada Gambar 3.



Gambar 3. Tahap akhir proses evaluasi

Gunung Djati Conference Series
ISSN 2774-6585
Volume 3 (2021)

3 Hasil dan Pembahasan

Dalam penelitian ini, kami menggunakan data yang merupakan data sekunder yang berupa data berita dari BBC News. Pengolahan data ini diimplementasikan menggunakan Bahasa python perbandingan data pada tahapan preprosesing dan sebelum dilakukan preprocessing dapat dilihat pada Gambar 4.

	category	text	text_original
0	tech	[tv, future, in, the, hands, of, viewers, with...	tv future in the hands of viewers with home th...
1	business	[worldcom, boss, left, books, alone, former, w...	worldcom boss left books alone former worldc...
2	sport	[tigers, wary, of, farrell, gamble, leicester,...	tigers wary of farrell gamble leicester say ...
3	sport	[yeading, face, newcastle, in, fa, cup, premie...	yeading face newcastle in fa cup premiership s...
4	entertainment	[ocean, s, twelve, raids, box, office, ocean, ...	ocean s twelve raids box office ocean s twelve...

Gambar 4. Perbandingan data tahap pre-processing dengan data asli

Tapahap encoding ini merubah bentuk teks menjadi numerik. Hal ini silakukan sebelum proses training menggunakan algoritma SVM. tampilan data setelah dilakukan proses encoding ditunjukkan pada Gambar 5.

```
{'tv': 4648, 'future': 1835, 'hand': 1996, 'viewer': 4790, 'home': 2095,
'theatre': 4487, 'system': 4407, 'plasma': 3322, 'digital': 1238, 'video':
4786, 'recorder': 3631, 'move': 2907, 'living': 2608, 'room': 3830, 'way':
4859, 'people': 3249, 'watch': 4855, 'different': 1234, 'five': 1724,
'year': 4979, 'time': 4526, 'accord': 31, 'expert': 1581, 'panel': 3187,
'gather': 1861, 'annual': 200, 'consumer': 954, 'electronics': 1427, 'show':
4048, 'la': 2485, 'discuss': 1266, 'new': 2992, 'technology': 4445,
'impact': 2181, 'one': 3090, 'favorite': 1652, 'lead': 2524, 'trend': 4607,
'programmer': 3476, 'content': 961, 'deliver': 1165, 'via': 4778, 'network':
2987, 'cable': 622, 'satellite': 3898, 'telecom': 4450, 'company': 869,
'broadband': 571, 'service': 3993, 'provider': 3514, 'front': 1814,
'portable': 3366, 'device': 1225, 'ce': 694, 'personal': 3269, 'pvr': 3539,
'box': 532, 'like': 2588, 'tivo': 4532, … }
```

Gambar 5. Hasil proses encoding

Implementasi Algoritma SVM ini menggunakan library yang sudah disediakan dengan source code yang ditampilkan pada Gambar 6.

```
[ ]  SVM = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
     SVM.fit(Train_X_Tfidf,Train_Y)
     predictions_SVM = SVM.predict(Test_X_Tfidf)
     print("SVM Accuracy Score -> ",accuracy_score(predictions_SVM, Test_Y)*100)
```

Gambar 6. Source code implementasi algoritma SVM

Hasil klasifikasi diperoleh nilai akurasi mencapai 98.35 % dengan rata-rata presisi 90%, recall 98%, F1-Score 98% dan Support sebesar 668. Hasil implementasi ditampilkan pada Gambar 7.

	precision	recall	f1-score	support
0	0.97	0.96	0.97	161
1	0.98	1.00	0.99	120
2	0.98	0.98	0.98	125
3	0.99	0.99	0.99	136
4	0.99	0.99	0.99	126
accuracy			0.98	668
macro avg	0.98	0.98	0.98	668
weighted avg	0.98	0.98	0.98	668

Gambar 7. Hasil implementasi klasifikasi

4 Simpulan

Penelitian ini dilakukan dikarenakan dalam membagikan sebuah berita pengklasifikasian teks belum dikategorikan pada konten berita digital. Pengklasifikasi menggunakan teknik Support Vector Machines (SVM). Sumber data yang digunakan dalam penelitian ini yaitu data sekunder berupa data berita dari BBC News yang telah diberikan kategori atau label. Label ini terdiri atas Teknologi, bisnis, sport, entertaimen, dan politik. Yang kemudian dataset ini digunakan untuk training terhadap system agar dapat memprediksi berita-berita lainnya. Terdiri dari 5 kategori berita dan tahapan preprocessing terhadap dataset ini terbagi menjadi 5 tahapan yaitu Removing blankspace, Casefolding, Tokenizing, Encoding. Hasil eksperimen yang dilakukan untuk memecahkan masalah klasifikasi konten berita digital, dapat disimpulkan bahwa terdapat 2224 dataset dan 5 kategori dengan 70% data yang ditraining dan 30% data yang ditesting.hasil eksperimen menggunakan metode Support Vector Machine diperoleh nilai akurasi mencapai 98.35 % dengan rata-rata presisi 90%, recall 98%, F1-Score 98% dan Support sebesar 668.

Referensi

[1] D. A. Pisner and D. M. Schnyer, “Support vector machine,” in *Machine Learning: Methods and Applications to Brain Disorders*, 2019, pp. 101–121.
[2] A. Mahinovs, A. Tiwari, R. Roy, and D. Baxter, *Text classification method review*. 2007.
[3] F. Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM Computing Surveys*, vol. 34, no. 1. pp. 1–47, 2002, doi: 10.1145/505282.505283.
[4] S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, “Particle swarm optimization for parameter determination and feature selection of support vector machines,” *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1817–1824, 2008, doi: 10.1016/j.eswa.2007.08.088.
[5] O. Maimon and L. Rokach, *Soft computing for knowledge discovery and data mining*. 2008.
[6] G. Tsoumakas, I. Katakis, and I. Vlahavas, *Data Mining and Knowledge Discovery Handbook Second Edition*. 2010.
[7] F. Gorunescu, *Data Mining: Concepts, models and techniques*. Springer, 2011.
[8] F. Gorunescu, “Data Mining Techniques and Models,” in *Data Mining*, 2011, pp. 185–317.
[9] M. W. Berry and J. Kogan, *Text Mining: Applications and Theory*. 2010.
[10] G. Krishnalal, S. B. Rengarajan, and K. G. Srinivasagan, “A New Text Mining Approach Based on HMM-SVM for Web News Classification,” *Int. J. Comput. Appl.*, vol. 1, no. 19, pp. 103–109, 2010, doi: 10.5120/395-589.
[11] A. Nurhadi, “Klasifikasi Konten Berita Digital Bahasa Indonesia Menggunakan Support Vector Machines (SVM) Berbasis Particle Swarm Optimization (PSO),” *J. Bianglala Inform.*, vol. 3, no. 2, pp. 1–9, 2015.
[12] S. Alsaleem, “Automated Arabic Text Categorization Using SVM and NB.,” *Int. Arab. J. e*

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

Technol., vol. 2, no. 2, pp. 124–128, 2011.

[13] Jumadi, D. S. Maylawati, B. Subaeki, and T. Ridwan, "Opinion mining on Twitter microblogging using Support Vector Machine: Public opinion about State Islamic University of Bandung," in *Proceedings of 2016 4th International Conference on Cyber and IT Service Management, CITSM 2016*, 2016, doi: 10.1109/CITSM.2016.7577569.

[14] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2006.