Tugas Proyek Akhir Praktikum Mata Kuliah Information Retrieval

“Perbandingan Performa Metode Klasifikasi Berita Online Berbahasa Inggris”

Kelompok 9:

Muhamad Feriyanto (222011347)

Muhammad Ruza F S (222011587)

Imam Sujono (222011830)

3SD1

Politeknik Statistika STIS

2022

Jakarta, 25 Desember 2022	Jakarta, 25 Desember 2022	Jakarta, 25 Desember 2022
(Muhammad Ruza F S)	(Muhamad Feriyanto)	(Imam Sujono)

Perbandingan Performa Metode Klasifikasi Berita *Online* Berbahasa Inggris

ABSTRAK

Pada saat ini, berita *online* menjadi salah satu sumber informasi utama bagi masyarakat. Dalam membaca berita *online* ini, setiap orang mempunyai preferensinya masing-masing mengenai topik berita yang disukainya. BBC News sebagai salah satu platform berita terkemuka menyediakan berita *online* dalam berbagai kategori. Berita *online* ini seringkali ditampilkan pada halaman depan aplikasi lain (pihak kedua), misalnya *browser*. Pihak kedua perlu untuk mengklasifikasikan seluruh berita dari berbagai sumber yang masuk ke basis datanya. Oleh karena itu, mereka harus membangun sistem klasifikasi berita *online* yang andal. Selanjutnya, hasil klasifikasi ini digunakan sebagai acuan untuk memberikan saran berita *online* kepada pengguna sesuai preferensinya. Hal ini akan meningkatkan kepuasan pengalaman pengguna dalam menggunakan aplikasi penyedia saran berita *online*. Namun, algoritma klasifikasi teks mempunyai pendekatannya masing-masing sehingga memunculkan perbedaan performa dalam mengklasifikasikan berita *online*. Penelitian ini bertujuan untuk membandingkan performa algoritma Naive Bayes, Rocchio, k Nearest Neighbor, dan Support Vector Machine dalam mengklasifikasikan berita *online*. Data yang digunakan dalam penelitian ini adalah 2225 data sampel berita *online* dari BBC News periode 2004 sampai dengan 2005 dengan lima kategori, yaitu bisnis, hiburan, politik, olahraga, dan teknologi. Dari hasil penelitian ini, didapat bahwa metode klasifikasi berita *online* dengan algoritma Support Vector Machine mempunyai performa yang terbaik dengan nilai weighted F1-score sebesar 99 pesen, weighted precision sebesar 99 persen, dan weighted recall sebesar 99 persen.

Kata kunci*—klasifikasi, berita online, Naive Bayes, Rocchio, kNN, SVM*

PENDAHULUAN

Pada saat ini, berita *online* merupakan salah satu sumber informasi yang paling sering diakses oleh masyarakat. Berita *online* ini tidak hanya disediakan oleh perusahaan yang bergerak di bidang massa tetapi juga disediakan oleh entitas yang beragam. Entitas ini berperan sebagai pihak kedua yang menjembatani media massa dan pembaca, misalnya ketika kita menggunakan mesin peramban (*browser engine*). Pada halaman depan, banyak berita dengan beragam kategori telah ditampilkan ke pengguna. Bahkan hal ini terjadi ketika pengguna sama sekali belum mengetikkan kueri atau kata kunci terkait berita yang ingin dibaca.

Jumlah berita *online* terus bertambah seiring berjalannya waktu. Di sisi lain, pengguna membutuhkan saran berita yang sesuai dengan preferensinya. Untuk itu, pihak kedua penyedia akses ke berita *online* perlu mengkategorikan berita yang masuk di basis datanya ke dalam kategori-kategori yang sesuai. Kategori ini nantinya akan digunakan untuk pertimbangan dalam memberikan saran berita *online* kepada pembaca.

BBC News merupakan salah satu media massa terkemuka yang menyediakan beragam berita *online* di situsnya. Berita ini tersedia dalam berbagai bahasa. Namun, penelitian kali ini menggunakan data berita *online* BBC News berbahasa Inggris. Salah satu pertimbanganya, yaitu berita *online* berbahasa Inggris menjadi berita *online* yang paling sering dibaca oleh pelanggan BBC News.

Berita-berita ini seringkali akan digunakan oleh pihak kedua untuk memberikan saran berita *online* kepada pengguna aplikasinya. Kategori berita yang ada di BBS News dapat berubah sewaktu-waktu sehingga pihak kedua perlu membangun sendiri algoritma pengklasifikasian berita.

Proses pengklasifikasian berita termasuk ke dalam klasifikasi teks. Metode klasifikasi yang sering digunakan antara lain Naive Bayes, Rocchio, k Nearest Neighbors (kNN), dan Support Vector Machine (SVM). Masing metode mempunyai pendekatannya sendiri sehingga menghasilkan performa yang berbeda-beda dalam klasifikasi data dengan format teks, termasuk berita *online*.

Penggunaan algoritma klasifikasi berita *online* yang mempunyai performa paling baik merupakan hal yang penting. Penggunaan algoritma yang paling tepat dapat berdampak positif terhadap layanan informasi berita yang disediakan oleh pihak kedua. Ketika berita-berita yang disarankan untuk dibaca sesuai dengan minat (preferensi) pembaca maka pengguna akan lebih sering menggunakan jasanya.

Dengan demikian, diperlukan suatu penelitian yang bertujuan untuk melakukan klasifikasi berita *online* menggunakan keempat metode tersebut dan kemudian membandingkan performanya. Hasil penelitian ini nantinya dapat digunakan sebagai bahan pertimbangan bagi beberapa pihak ketika ingin membangun sistem pengklasifikasian berita *online*.

METODOLOGI

Data yang digunakan pada proyek ini adalah berita bahasa Inggris yang bersumber dari BBC pada periode 2004 - 2005 dengan total 2225 data. Topik yang digunakan pada keseluruhan berita berjumlah 5 topik, yakni teknologi (tech), politik (politics), olahraga (sports), hiburan (entertainment), dan bisnis (business). Setiap dokumen memiliki satu topik yang telah pelabelan topiknya dilakukan dengan manual.

Sebelum dilakukan pembentukan model klasifikasi berita, dilakukan *preprocessing* untuk mendapatkan dokumen yang bersih. Tahapan *preprocessing* yang dilakukan pada proyek kali ini di antaranya adalah lower-case, penghapusan karakter yang bukan alfabet seperti tanda hubung (-), angka, spasi yang berlebih, dan lain sebagainya, lematisasi, dan menghapus kata-kata *stopwords*.

Pembentukan model untuk klasifikasi berita dari BBC yang akan dilakukan, digunakan beberapa model klasifikasi yang nantinya akan dibandingkan untuk mendapatkan model klasifikasi terbaik. Metode yang digunakan untuk membentuk model diantaranya yakni, Naive Bayes, Rocchio classification, K- Nearest Neighbor (KNN), dan Support Vector Machine (SVM).

Naive Bayes

Metode Naive Bayes yang digunakan adalah multinomial Naive Bayes yang merupakan sebuah metode pembelajaran berdasarkan peluang (*probabilistic learning method*). Menurut Manning (2008), peluang dari dokumen d dikelompokkan ke dalam kelas c dapat dituliskan dalam rumus:

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

dengan

- $P(t_k|c)$ adalah conditional probability dari term t_k yang ada dalam kelas c
- $P(c)$ adalah prior probability dari c

Rocchio

Rocchio merupakan algoritma yang merepresentasikan data dalam ruang vektor dan membagi ruang vektor tersebut menjadi beberapa bagian berdasarkan centroid yang ada. Menurut Manning et al.,(2008), Algoritma Rocchio mudah dan sederhana untuk digunakan, tetapi akurasinya buruk ketika jarak antar centroid sama. Centroid adalah pusat dari setiap

kelas dalam ruang vektor. Nilai centroid ini diperoleh dengan menghitung jarak rata-rata dari setiap dokumen. Centroid dapat dihitung dengan persamaan

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \epsilon D_c} \vec{v}(d)$$

dengan

- Dc adalah sekumpulan dokumen yang termasuk kelas c
- v(d) adalah vektor kata-kata di kelas c

K-Nearest Neighbor

K-Nearest Neighbor (k-NN) adalah salah satu metode proses pengelompokkan data ke dalam kelas yang sudah ditentukan di awal berdasarkan jarak terkecil/terdekat. Berikut merupakan langkah-langkah dalam menggunakan metode k-NN:

- Tentukan nilai K, dalam penelitian kali ini menggunakan nilai K=5
- Menghitung jarak antara data yang baru dari setiap label data
- Tentukan data yang memiliki jarak yang paling kecil
- Klasifikasikan data yang baru ke label data yang sebagian besar K-Nearest Neighbour dipilih dari metrik jarak

Support Vector Machine

Menurut Manning, et all., (2008), Support Vector Machine (SVM) merupakan salah satu jenis dari metode pengklasifian margin yang besar (*large margin classifier*), yang berarti bahwa ruang vektor yang berdasarkan metode *machine learning* yang memiliki tujuan untuk menemukan batasan (*decision boundary*) di antara dua kelas (kategori) yang memaksimalkan jarak dari sembarang titik pada data training (terdapat kemungkinan untuk mengabaikan beberapa titik sebagai *outlier* atau *noise*). SVM mungkin tidak lebih baik dari metode *machine learning* lainnya (kecuali dalam situasi *data training* yang sedikit), tapi SVM merupakan metode yang canggih dan memberikan daya tarik dalam segi teoritis dan empiris.

SVM secara khusus mendefinisikan kriteria yang dicari untuk permukaan keputusan (*decision surface*) yang memaksimalkan jarak dari semua titik data yang ada. Jarak ini dari permukaan keputusan (*decision surface*) ke titik yang paling dekat dengan data menentukan margin dari *classifier*. Metode ini berarti bahwa fungsi keputusan (*decision function*) untuk SVM secara keseluruhan merupakan spesifikasi biasanya ditentukan oleh subset dari data yang didefinisikan sebagai pemisah yang berada pada paling luar dari kelompok tersebut. Titik-titik ini dalam SVM disebut sebagai *support vector*.

HASIL DAN PEMBAHASAN

Analisis Deskriptif

Pada dataset yang digunakan, terdapat 329 berita berkategori *politics*, 412 berita berkategori *business*, 324 berita berkategori *tech,* 413 berita berkategori *sport*, dan 302 berita berkategori *entertainment*. Berdasarkan *worldcloud* yang terbentuk untuk masing-masing kategori berita, terdapat kata-kata yang lebih sering muncul dibandingkan kata lainnya. Kata-kata ini merupakan kata yang menjadi ciri khas dari suatu kategori berita *online*.

Tabel 1. Komponen Penyusun Berita

Kategori	Rata-rata Kata	Rata-rata Karakter	Rata-rata Kata Unik
Politics	447,60	2607,16	234,63
Business	334,42	1986,23	192,18
Tech	510,88	2994,19	262,71
Sport	339,06	1911,50	191,70
Entertainment	336,72	1923,19	189,57

Dari tabel di atas, terlihat bahwa berita dengan rata-rata kata terbanyak adalah berita *online* berkategori *tech* sedangkan berita dengan rata-rata kata paling sedikit adalah berita *online* berkategori *business*.

Setiap kategori berita *online* mempunyai kata-kata khasnya masing-masing. Pertama, kategori *business* didominasi oleh kata *company*, *share*, *firm*, *sale*, *market*, *growth*, dan *price*. Kedua, kategori *entertainment* didominasi oleh kata *film*, *best*, *award*, *star*, *show*, *music*, *actor* dan *actress* . Ketiga, kategori *politics* didominasi oleh kata *labour*, *election*, *party*, *government*, *minister*, Blair, *tory*, dan *people*. Keempat, kategori *sport* didominasi oleh kata *game*, *player*, *england*, *win*, *final*, *match*, dan *time* . Kelima, kategori *tech* didominasi oleh kata *game*, *mobile*, *people*, *technology*, *computer*, *music*, *digital*, dan *software*.

Naive Bayes

Berdasarkan pembentukan model yang dilakukan dengan metode Multinomial Naive Bayes, didapatkan nilai precision, recall, dan f1-score pada masing-masing topik sebagai berikut.

Tabel 2. Evaluasi Performa Metode Naive Bayes

Topik	Precision	Recall	F1-Score
Business	0,98	0,96	0,97
Entertainment	1,00	0,83	0,91
Politics	0,92	0,99	0,95
Sport	0,92	1,00	0,96
Tech	0,95	0,96	0,95

Berdasarkan tabel di atas, didapatkan nilai precision yang tidak jauh berbeda untuk setiap topik yang ada, nilai recall kecenderungan tinggi (lebih dari 80%), dan nilai F1-Score yang tidak jauh berbeda. Nilai precision, recall, dan F1-Score yang dihasilkan telah mampu merepresentasikan model dengan sangat baik.Sedangkan secara keseluruhan, nilai F1-Score yang didapatkan sebesar 0,95. Sehingga, sebesar 95% klasifikasi yang dibentuk pada model Naive Bayes sudah tepat sesuai dengan topik yang ada pada data training.

Rocchio

Berdasarkan pembentukan model yang dilakukan dengan metode Rocchio, didapatkan nilai precision, recall, dan F1-score pada masing-masing topik sebagai berikut.

Tabel 3. Evaluasi Performa Metode Rocchio

Topik	Precision	Recall	F1- Score
Business	0,93	0,98	0,96
Entertainment	0,96	0,95	0,96
Politics	0,98	0,91	0,94
Sport	1,00	0,99	0,99
Tech	0,93	0,96	0,94

Dapat kita lihat pada tabel diatas, didapatkan nilai precision yang tidak jauh berbeda untuk setiap topik yang ada, nilai recall yang tinggi yaitu lebih dari 90% untuk setiap topik, dan nilai F1-Score yang tidak jauh berbeda. Nilai precision, recall, dan F1-Score yang dihasilkan telah mampu merepresentasikan model dengan sangat baik.Sedangkan secara keseluruhan, nilai F1-Score yang didapatkan sebesar 0,9596. Sehingga, sebesar 95,96% klasifikasi yang dibentuk dengan metode Rocchio sudah tepat sesuai dengan topik yang ada pada data training.

K-Nearest Neighbor

Berdasarkan pembentukan model yang dilakukan dengan metode K-Nearest Neighbor (k-NN), didapatkan nilai precision, recall, dan F1-score pada masing-masing topik sebagai berikut.

Tabel 4. Evaluasi Performa Metode K-Nearest Neighbor

Topik	Precision	Recall	F1-Score
Business	0,97	0,90	0,93
Entertainment	0,99	0,88	0,93
Politics	0,87	0,97	0,91
Sport	0,98	0,98	0,98

Tech	0,92	0,99	0,95

Dari tabel diatas, bisa kita lihat bahwa nilai precision, recall, dan F1-Score untuk setiap topik memiliki nilai yang tinggi. Dengan menggunakan metode k-NN nilai precision, recall, dan F1-Score yang dihasilkan telah mampu merepresentasikan model dengan sangat baik.Sedangkan secara keseluruhan, nilai F1-Score yang didapatkan sebesar 0,9416. Sehingga, sebesar 94,916% klasifikasi yang dibentuk dengan metode K-Nearest Neighbor sudah tepat sesuai dengan topik yang ada pada data training.

Support Vector Machine

Berdasarkan pembentukan model yang dilakukan dengan metode Support Vector Machine (SVM), didapatkan nilai precision, recall, dan F1-score pada masing-masing kategori sebagai berikut.

Tabel 5. Evaluasi Performa Metode SVM

Topik	Precision	Recall	F1-Score
Business	0,99	0,98	0,98
Entertainment	1,00	1,00	1,00
Politics	0,98	0,99	0,98
Sport	1,00	1,00	1,00
Tech	0,99	0,99	0,99

Dari hasil tersebut, nilai precision, recall, dan F1-score tidak berbeda jauh antar kelas. Hal ini mengindikasikan bahwa model klasifikasi yang dibentuk menggunakan metode SVM memiliki performa yang stabil dalam memprediksi kategori dari berita *online*. Misalnya, untuk kategori *business* nilai precisionnya sebesar 0,99 berarti 99 persen hasil prediksi bahwa berita *online* yang dimaksud termasuk dalam kategori *business* adalah benar. Kemudian, nilai recallnya sebesar 0,98, berarti dari seluruh berita *online* yang termasuk kelas business, 98 persen berhasil diprediksi dengan tepat.

Untuk melihat performa model secara keseluruhan dapat menggunakan average F1-score. Nilai average F1-score yang didapat sebesar 0,9910 yang berarti model sangat baik dalam mengklasifikasikan berita *online*.

Perbandingan Performa

Perbandingan nilai F1-Score dilakukan untuk membandingkan metode dengan model terbaik untuk mengklasifikasikan dokumen berita tersebut. Semakin tinggi nilai F1-Score, maka semakin baik model. Berikut tabel perbandingan F1-Score setiap metode.

Tabel 6. Perbandingan Performa

Metode	F1-Score
Naive Bayes	0,9506
Rocchio Classification	0,9596
KNN	0,9416
SVM	0,9910

Nilai F1-score di atas merupakan nilai weighted F1-score. Weighted Average F1-score dipilih karena beberapa alasan. Pertama, jumlah sampel berita *online* untuk masing-masing kategori tidak sama dengan standar deviasi yang cukup besar, yaitu 52,57. Selain itu, jika kita melihat berita-berita *online* yang terdapat di platform BBC News, terdapat beberapa kategori yang cukup mendominasi dalam hal jumlah berita *online*. Oleh karena itu, penggunaan weighted Average F1-score akan memberikan penimbang yang lebih besar kepada kategori dengan jumlah berita *online* yang lebih banyak. Dengan demikian, nilai average F1-score yang digunakan sebagai ukuran evaluasi performa model lebih representatif.

Berdasarkan tabel tersebut, keempat model mempunyai performa yang sangat baik dalam mengklasifikasikan berita *online* ke dalam tiap-tiap kategori. Kemudian, didapatkan nilai weighted average F1-Score tertinggi dengan menggunakan metode SVM (C = 1, kernel = linear, degree = 3, gamma = auto), yakni sebesar 99,10%. Dengan demikian, dapat dikatakan bahwa model klasifikasi yang dibentuk menggunakan metode SVM paling cocok untuk mengklasifikasikan kategori dari suatu berita *online* (dalam studi kasus ini menggunakan berita *online* BBC News).

KESIMPULAN DAN SARAN

Berdasarkan penerapan metode Naive Bayes, Rocchio Classification, kNN, dan SVM pada data berita *online* yang bersumber dari BBC News periode tahun 2004-2005, metode SVM (C = 1, kernel = linear, degree = 3, gamma = auto) memberikan nilai evaluasi model seperti *precision, recall,* dan F1-Score yang relatif paling tinggi dari metode lainnya. Hal ini dapat disimpulkan bahwa pemodelan klasifikasi berita dengan menggunakan metode SVM

(C = 1, kernel = linear, degree = 3, gamma = auto) merupakan pemodelan klasifikasi berita yang terbaik. Dengan menggunakan prinsip *parsimony,* berdasarkan hasil evaluasi dari metode yang digunakan, nilai F1-Score dari metode terbaik (SVM) tidak jauh berbeda dengan nilai F1-Score dari metode Naive Bayes. Sehingga, jika ingin menggunakan model klasifikasi dengan metode yang simpel maka metode Naive Bayes masih dapat digunakan untuk mengklasifikasikan topik berita dengan baik.

Keterbatasan yang didapatkan pada proyek ini yakni satu dokumen hanya memiliki satu topik berikutnya. Sehingga, untuk penelitian selanjutnya, dapat digunakan klasifikasi berita lebih dari satu topik karena pada realitanya, kategori dari berita tidak terbatas oleh satu kategori.

LAMPIRAN

- Link google collaboratory : https://colab.research.google.com/drive/1zZ1i4SFUQZ8stVcC-OBjQ0vmI7TSqulg?usp=share_link
- Link dataset : https://raw.githubusercontent.com/cigdemtuncer/NewsTextClassification/main/bbc-text.csv
- Link keterangan dataset : https://medium.com/analytics-vidhya/bbc-news-text-classification-a1b2a61af903
- Pembagian Tugas

No	Langkah	Tugas
Coding		
1	Preprocessing	Muhamad Feriyanto
2	Analisis Deskriptif	Imam Sujono
3	Naive Bayes	Muhamad Feriyanto
4	Rocchio Classification	Muhamma Ruza F.S
5	kNN	Muhamma Ruza F.S

6	SVM	Imam Sujono
7	Perbandingan Evaluasi	Muhamad Feriyanto
Laporan		
1	Abstrak	Imam Sujono
2	Pendahuluan	Imam Sujono
3	Metodologi	Muhamad Feriyanto dan Muhammad Ruza F.S
4	Hasil dan Kesimpulan	- Analisis Deskriptif : Imam Sujono - Naive Bayes : Muhamad Feriyanto - Rocchio : Muhammad Ruza F.S - kNN : Muhammad Ruza F.S - SVM : Imam Sujono - Perbandingan : Muhamad Feriyanto
5	Penutup (Kesimpulan dan Saran)	Muhamad Feriyanto