ISSN: 1907-7912 EISSN: 2622-8092

Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Konten Berita Olahraga

Ari Putra Wibowo [(1)] Arief Soma Darmawan [(2)]
STMIK Widya Pratama Pekalongan
Jl. Patriot 25 Pekalongan Telp (0285) 427816
[(1)] email: ariputra.stmikwp@gmail.com
[(2)] email: soma98980@yahoo.com

ABSTRAK

Penelitian terkait klasifikasi telah dipelajari secara luas untuk keperluan data mining, machine learning dan database serta information retrieval yang diaplikasikan untuk menentukan target pemasaran, diagnosis medis, konten berita serta klasifikasi dokumen. Klasifikasi teks menjadi pembahasan yang ramai dibahas oleh peneliti selama dua dekade terakhir. Meskipun didalam metode dan tekniknya selalu ada pembaharuan namun kebutuhannya masih terus berkembang dan tidak pernah berakhir. Kemampuan untuk melakukan klasifikasi dokumen ke dalam kategori tertentu sangat membantu untuk menghadapi informasi yang berlebihan. Klasifikasi dokumen teks secara otomatis dikembangkan karena pekerjaan manual tidak lagi efektif. Pada penelitian ini akan dibahas bagaimana algoritma naïve bayes diterapkan untuk mengklasifikasi konten berita olahraga. Naïve bayes merupakan salah satu algoritma klasifikasi berbasis peluang. Maka dari itu, naïve bayes akan menghitung probabilitas kemunculan kata yang mempresentasikan dokumen teks dari konten berita. Berdasarkan penelitian yang dilakukan diperoleh hasil rata-rata akurasi adalah 69,27%. Dengan nilai akurasi terbesar yaitu 75,00% dengan data dokumen teks yang digunakan sebanyak 20% dari keseluruhan dokumen teks..

Kata Kunci: **text mining, naïve bayes, klasifikasi konten berita**

1. PENDAHULUAN

Penelitian terkait klasifikasi telah dipelajari secara luas untuk keperluan data mining, machine learning dan database serta information retrieval yang diaplikasikan untuk menentukan target pemasaran, diagnosis medis, konten berita serta klasifikasi dokumen (Aggarwal and Zhai 2013). Klasifikasi teks menjadi pembahasan yang ramai dibahas oleh peneliti selama dua dekade terakhir. Meskipun didalam metode dan tekniknya selalu ada pembaharuan namun kebutuhannya masih terus berkembang dan tidak pernah berakhir (Zakzouk and Mathkour 2012). Telah banyak penelitian yang dilakukan untuk membandingkan berbagai pengklasifikasian berbasis machine learning (Y. Yang and Liu 1999) (Colas and Brazdil 2006).

Seiring berkembangnya teknologi terutama di bidang ilmu komputer, memungkinkan untuk mengakses informasi kapan saja dan di mana saja (Wongso et al. 2017). Dengan banyaknya jenis informasi yang tersedia, pengelompokan informasi menjadi tugas yang menantang agar informasi yang ada lebih mudah dipahami (C. C. Yang, Chen, and Hong 2003). Kemampuan untuk melakukan klasifikasi dokumen ke dalam kategori tertentu sangat membantu untuk menghadapi informasi yang berlebihan. Klasifikasi dokumen teks secara otomatis dikembangkan karena pekerjaan manual tidak lagi efektif (Wong and Abednego 2015).

Beberapa penelitian terkait klasifikasi dokumen teks antara lain penelitian yang dilakukan oleh (Wijaya and Santoso 2016) melakukan klasifikasi dokumen untuk mengidentifikasi konten berita politik dan ekonomi menggunakan algoritma naïve bayes classification. Klasifikasi dokumen yang dilakukan adalah konten berbasaha Indonesia. Penelitian berikutnya yang dilakukan oleh (Anita et al. 2015) mengklasifikasi artikel untuk kategori kesehatan dengan menggunakan algoritma naïve bayes classification. Selanjutnya pada penelitian ini akan dilakukan klasifikasi dokumen teks kategori olahraga dengan konten

sepakbola dengan menggunakan naïve bayes classification. Sepakbola dipilih karena memang merupakan salah satu cabang olahraga yang memang sangat familiar dan digemari oleh berbagai kalangan masyarakat di Indonesia. Dokumen teks yang digunakan diperoleh dari portal berita olahraga https://www.bola.net pada penelitian ini akan dibahas bagaimana algoritma naïve bayes diterapkan untuk mengklasifikasi konten berita olahraga. Naïve bayes merupakan salah satu algoritma klasifikasi berbasis peluang.

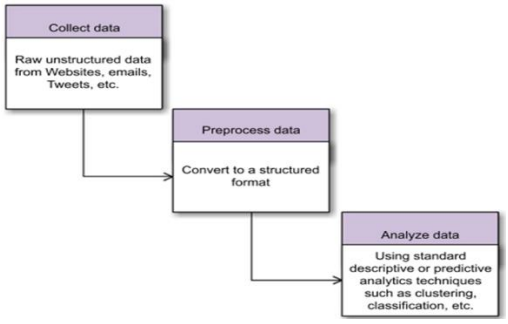
Maka dari itu, naïve bayes akan menghitung probabilitas kemunculan kata yang mempresentasikan dokumen teks dari konten berita yang digunakan (Wibowo and Jumiati 2017).

2. LANDASAN TEORI

2.1. *Text Mining*

Text mining atau yang juga biasa disebut text analytics dapat secara luas didefinisikan sebagai proses mengekstraksi informasi yang berguna dari berbagai macam dokumen melalui tahap identifikasi dan eksplorasi pola-pola menarik dalam data tekstual terstruktur dari berbagai jenis dokumen (Votano, Parham, and Hall 2004) seperti buku, halaman web, email , laporan atau deskripsi produk. Ada beberapa domain yang umum dilakukan text mining meliputi: mesin pencarian (search engine), melakukan ketegori teks ke satu atau beberapa kategori (text categoritation), mengelompokkan teks yang sejenis (text clustering), menemukan topik/tema dari suatu diskusi (concept/entity extraction), menemukan opini dari teks (sentiment analysis) serta meringkas dokumen, dan hubungan pembelajaran antara entitas yang dijelaskan dalam teks (entity relation modelling) (Truyens and Van Eecke 2014).

Tahapan dalam text mining dimulai dengan mengkonversi dokumen teks menjadi data semi-terstruktur. Setelah mengubah dokumen teks yang tidak terstruktur menjadi data semi-terstruktur, berikutnya dilakukan analisis data baik dengan tujuan mengklasifikasikan, mengelompokkan ataupun memprediksi. Teks yang tidak terstruktur perlu dikonversi menjadi dataset semi-terstruktur sehingga dapat menemukan pola dan lebih baik lagi, melatih model untuk mengetahui pola dalam teks baru dan teks yang kurang jelas(Witten 2004). Gambar 2.1 berikut ini menjelaskan tahapan dalam proses text mining.



Gambar 1 Tahapan Proses *Text Mining* (Witten 2004)

2.2. *Text Preprocessing*

Salah satu tantangan text mining adalah mengubah teks terstruktur dan semi terstruktur menjadi model ruang vektor terstruktur. Ini harus dilakukan sebelum melakukan text mining atau melakukan analisis. Langkah-langkah text preprocessing meliputi (Model et al. 2012):

1. Memilih ruang lingkup/domain yang akan diproses (buku, web, email dll).
2. Tokenize: memecah teks menjadi kata-kata tersendiri yang disebut token.
3. Remove stopwords (“stoping”): menghilangkan/menghapus kata konjungsi seperti kata: di, yang, ke dll.
4. Stem: menghapus awalan dan akhiran pada kata untuk memperoleh kata dasar dari kata yang digunakan.
5. Normalize spelling: menyatukan kesalahan ejaan dan variasi ejaan lainnya menjadi satu kata (token).
6. Detect sentence boundaries: menandai/memberi label dari setiap kalimat
7. Normalize case: Konversi teks menjadi semua huruf kecil atau semua huruf besar.

2.3. Naïve Bayes Classification

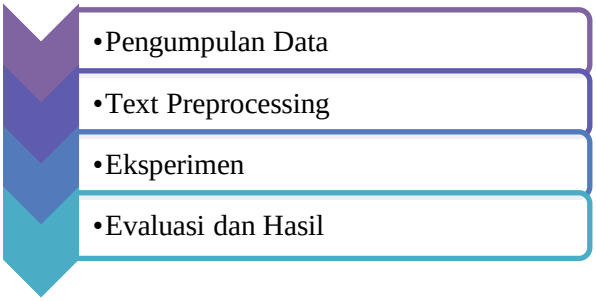
Naïve bayes classifiers adalah salah satu algoritma pengklasifikasi linear yang sederhana dengan kinerja yang sangat efisien. Penentuan problabilistik pada pengklasifikasian naïve bayes didasarkan pada teorema bayes, naive bayes terkenal dengan asumsi independen yaitu bahwa fitur-fitur dalam dataset sama-sama independen (Raschka 2014). Naive bayes classification menjadi salah satu algoritma yang banyak digunakan dalam melakukan klasifikasi dokumen (Wahyu 2014). Dalam prosesnya naive bayes clasification dibagi menjadi tahap training dan tahap testing (klasifikasi). Pada tahap training dilakukan analisis terhadap dokumen berupa pemilihan kosa kata, kata-kata yang terdapat dalam dokumen bisa saja menjadi representasi dokumen. Tahap berikutnya atau tahap testing adalah penentuan probabilitas untuk tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi (Hamzah 2012). Untuk menentukan probabilitas dari masing-masing term dapat menggunakan persamaan seperti berikut ini:

$$P(w_i|C) = \frac{count\ (w_i.C) + 1}{count\ (C) + |v|}$$

Dimana P(w_i |C) adalah probalilitas dari jumlah kata w_i pada kelas C, count (C) adalah jumlah kata di kelas C, |v| adalah jumlah kosa kata (vocabulary).

3. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian eksperimental, dalam penelitian ini akan dilakukan pengujian terhadap dokumen teks untuk mengetahui nilai akurasi terbaik dari beberapa percobaan yang dilakukan. adapun tahapan penelitian ini seperti pada Gambar 2 yang di dalamnya berisi tentang tahapan yaitu:



Gambar 2 Tahapan Penelitian

3.1. Pengumpulan Data

Data yang digunakan pada penelitian ini diperoleh dari situs berita orahraga https://www.bola.net/. Data berupa dokumen teks (judul artikel) terkait kategori dari konten berita olahraga khususnya sepakbola yang dimuat pada situs tersebut. Kategori konten berita olahraga yang digunakan pada penelitan ini adalah sebakbola Eropa (Liga Inggris, Liga Italia, Liga Spanyol) dan sepakbola Indonesia (Liga 1 Indonesia).

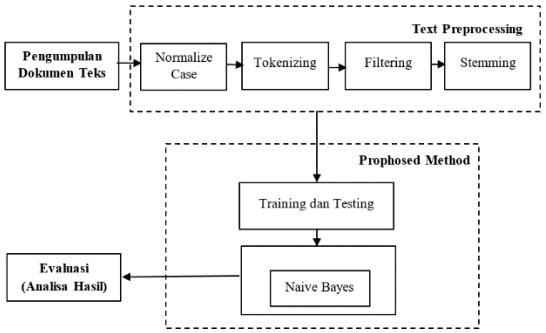
3.2. Text Preprocessing

Setelah dokumen teks telah diperoleh, tahap selanjutnya adalah text preprocesing yang meliputi :

1) Normalize Case
2) Tokenizing
3) Filtering
4) Stemming

3.3. Eksperimen

Untuk melakukan pengujian model, dilakukan dengan menggunakan Rapid Miner, dari algoritma yang sudah ditentukan maka selanjutnya dataset yang sudah ada akan diolah sehingga menghasilkan model yang akan diinginkan.



Gambar 3 Kerangka Pemikiran

3.4. Evaluasi dan Hasil

Setelah melakukan eksperimen terhadap semua dokumen teks dengan model yang diusulkan maka akan menghasilkan nilai-nilai akurasi dari model yang digunakan kemudian hasil tersebut dianalisa dan dievaluasi. Dari hasil evaluasi selanjutnya dapat ditarik kesimpulan dari penelitian dan eksperimen ini.

4. HASIL DAN PEMBAHASAN

4.1. Pengumpulan Data

Dokumen teks yang digunakan pada penelitian ini diperoleh dari portal berita olahraga www.bola.net terbitan tanggal 10 sampai 15 november 2019. Ada empat kategori pada dokumen teks yaitu Liga Inggris, Liga Italia, Liga Spanyol dan Liga Indonesia, masing-masing kategori terdapat 100 dokumen teks.

Tabel 1 Dokumen teks liga Inggris

No	Dokumen teks Liga Inggris
1	Manchester United Sudah Tahu Kapan Bisa Memainkan Pogba
2	Bukan Daniel James, Inilah Pemain Tercepat Manchester United Musim Ini
3	Klopp Pastikan Liverpool Takkan Rekrut Zlatan Ibrahimovic Karena Dia Bekas Pemain MU
4	Pengakuan Henderson: Pernah Nyaris Habisi Luis Suarez
5	Jika Antarkan Liverpool Juara, Ini yang Akan Diterima Jurgen Klopp
6	Eks Liverpool Ini Berfantasi Kane Bisa Masuk Skuat Liverpool dan Singkirkan Firmino
7	Hanya Masalah Waktu Sebelum Kieran Tierney 'Meledak' di Arsenal
8	Olivier Giroud Minta Pertemuan dengan Chelsea, Ada Apa?

9	Fowler: Jangan Bicara Juara Dulu, Liverpool!
10	Mengapa Liverpool Tidak Kunjung Cari Pengganti Coutinho?
…	…
100	Sindiran Mourinho pada Kompany: Tak Tahu Terima Kasih

Tabel 2 Dokumen teks liga Italia

No	Dokumen teks Liga Italia
1	Kembali, Maurizio Sarri Disama-samakan dengan Josep Guardiola
2	Deulofeou CLBK dengan AC Milan Masih Mungkin Terjadi
3	AC Milan Terdepan untuk Rekrut Granit Xhaka
4	Carlo Ancelotti Pertimbangkan Balik ke AC Milan
5	Dihargai 300 Juta Euro oleh Presiden Brescia, Begini Respon Sandro Tonalli
6	Demi Pulangkan Ibrahimovic, Milan Siap Tumbalkan Tiga Pemain
7	Duet Maut! Pele Idamkan Lionel Messi Sebagai Duetnya di Lini Serang
8	Eks Juventus: Capello Benar, Ronaldo Sudah tak Bisa Lewati Lawan
9	Cara Juventus Selesaikan Masalah Ronaldo: Tanpa Sanksi, Hanya Tuntut Minta Maaf
10	Direktur Juventus di Manchester, Misi Bawa Pulang Paul Pogba?
…	…
100	Juventus vs AC Milan, Ronaldo dan De Ligt Siap Dimainkan?

Tabel 3. Dokumen teks liga Spanyol

No	Dokumen teks Liga Spanyol
1	Hubungan Griezmann dan Messi di Barcelona 'Sedang Dalam Proses'
2	Gerard Pique Mengaku Punya Nomor Ponsel Florentino Perez, Buat Apa?
3	Bagi Courtois, Clean Sheet tak Terlalu Penting
4	Bukan Inter Milan, Luka Modric Pilih Hijrah ke Amerika Serikat
5	Bagi Thibaut Courtois, Kritikan Dari Media Itu Tidak Penting
6	Kisah Lionel Messi: 'Leonel Mecci' yang Hampir Gabung Timnas Spanyol
7	3 Calon Pengganti Ivan Rakitic yang Dibidik Barcelona
8	Ronald Koeman Kemungkinan Jadi Pelatih Barcelona
9	Ivan Rakitic Curhat Sedang Sedih, Pelatih Barcelona Irit Komentar
10	Thibaut Courtois yang tak Pernah Rapuh dan Ambyar
…	…
100	7 Langkah Rodrygo Menjadi Bintang di Real Madrid

Tabel 4. Dokumen teks liga Indonesia

No	Dokumen teks Liga Indonesia
1	Diminta Mundur dari PSM, Ini Respon Darije Kalezic
2	Milan Petrovic Yakin Badak Lampung Bertahan di Liga 1
3	Arema Waspadai Ketajaman Lini Depan Persija
4	Lothar Matthaus Beber Efek Positif kesuksesan Indonesia Jadi Tuan Rumah Piala Dunia U-20
5	Milan Petrovic Senang Raihan Penggawa Badak Lampung Selama TC di Yogya
6	Arema Mulai Panasi Mesin Jelang Menjamu Persija Jakarta
7	Aji Santoso Akui Persebaya Kesulitan Hentikan Amido Balde
8	Gagalkan Penalti PSM, Persebaya Tak Salah Pasang Miswar Saputra
9	Pelatih Persija Tak Pikirkan Rekor Pertemuan Menghadapi Persela
10	Persebaya Syukuri Kemenangan Tipis dari PSM Makassar
…	…
100	Rendi Irwan Akui Kehadiran Aji Santoso Menambah Motivasi Pemain Persebaya

4.2. Text Preprocessing

Pada tahap teks preprocessing dilakukan beberapa tahapan untuk memperoleh kata-kata (bobot) yang mewakili setiap kategori.

4.2.1. Normalize Case

Pada normalize case dilakukan konversi teks pada dokumen, pada penelitian ini konversi teks diubah menjadi huruf kecil semua.

Misal :

Sebelum dilakukan normalize case :
Manchester United Sudah Tahu Kapan Bisa Memainkan Pogba

Setelah dilakukan normize case :
manchester united sudah tahu kapan bisa memainkan pogba

4.2.2. Tokenizing

Pada tahap tokenizing dilakukan pemecahan kalimat menjadi kata-kata sendiri.

Misal :

Sebelum dilakukan tokenizing :
manchester united sudah tahu kapan bisa memainkan pogba

Setelah dilakukan tokenizing :
manchester
united
sudah
tahu
kapan
bisa
memainkan
pogba

4.2.3. Filtering (Stoping)

Pada tahap filtering dilakukan proses menghilangkan/menghapus kata konjungsi

Tabel 5. Kamus kata stoping

adalah	bagai	cara	dari
adanya	bagaikan	caranya	daripada
adapun	bagaimana	cukup	datang
agak	bagaimanakah	cukupkah	dekat
amat	bagaimanapun	cukuplah	demi
antar	bagi	dahulu	demikian
antara	bagian	dalam	demikianlah
atau	bahkan	dan	dengan
...	...	...	...

4.2.4. Stemming

Pada tahap stemming dilakukan proses menghapus awalan dan akhiran pada kata untuk memperoleh kata dasar dari kata yang digunakan.

Tabel 6. Hasil stemming

Sebelum stemming	Setelah stemming
alasannya	alasan
dianggap	anggap
berbahaya	bahaya
terbaik	baik
pencetak	cetak
...	...

4.3. Eksperimen

Eksperimen pada penelitian ini menggunakan aplikasi rapid miner. Pada eksperimen ini dilakukan pengujian dengan pembagian dataset untuk menghitung tingkat akurasi dengan metode naïve bayes. Dari hasil percobaan diperoleh hasil seperti tabel 7 berikut :

Tabel 7. Hasil eksperimen

Prosentasi data training	Akurasi
0.1	62.50
0.2	60.00
0.3	68.33
0.4	72.50
0.5	75.00
0.6	69.58
0.7	70.36
0.8	69.69
0.9	72.78
1	72.00
Rata-rata	**69,27**

Dari penelitian yang sudah dilakukan dalam penelitian tentang klasifikasi berita olahraga sepakbola dengan 4 kategori dan setiap kategori memiliki 100 dokumen teks yang diperoleh dari portal berita olahraga www.bola.net diperoleh nilai akurasi rata-rata 69,27%. Hasil akurasi tertinggi didapat pada penggunaan 50 dokumen teks dengan nilai akurasi 75,00%. Sedangkan untuk hasil akurasi terendah didapat pada penggunaan 20 dokumen teks dengan nilai akurasi 60,00% .

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian dan eksperimen yang telah dilakukan untuk mengklasifikasi konten berita olahraga dengan menggunakan algoritma naïve bayes, diperoleh hasil yang baik dari model yang digunakan yaitu dengan rata-rata akurasi 69,27%.

5.2. Saran

Pada penelitian ini belum dilakukan pembobotan untuk dokumen teks yang digunakan, sehingga untuk penelitian selanjutnya bisa dilakukan pembobotan pada dokumen teks dengan

harapan untuk meningkatkan akurasi dari pengklasifikasian algoritma naïve bayes.

DAFTAR PUSTAKA

Aggarwal, Charu C., and Cheng Xiang Zhai. 2013. Mining Text Data. Mining Text Data. Vol. 9781461432. https://doi.org/10.1007/978-1-4614-3223-4.

Anita, Nur, Bagus Setya Rintyarna S T, M Kom, Lutfi Ali Muharom, S Si, and Sistem Bisnis Cerdas. 2015. "Klasifikasi Teks Dengan Naive Bayes Classifier Untuk Pengelompokan Teks Artikel," no. 1110651094.

Colas, Fabrice, and Pavel Brazdil. 2006. "Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks" 217: 169–78.

Hamzah, Amir. 2012. "Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis Amir." Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 3 (2011): 269–77. https://doi.org/1979-911X.

Model, The Generalized Vector-space, Preprocessing Text, Creating Vectors, and Processed Text. 2012. "Conceptual Foundations of Text Mining and Preprocessing Steps." Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, 43–51. https://doi.org/10.1016/B978-0-12-386979-1.00003-7.

Raschka, Sebastian. 2014. "Naive Bayes and Text Classification I - Introduction and Theory," 1–20. https://doi.org/10.13140/2.1.2018.3049.

Truyens, Maarten, and Patrick Van Eecke. 2014. "Legal Aspects of Text Mining." Computer Law and Security Review 30 (2): 153–70. https://doi.org/10.1016/j.clsr.2014.01.009.

Votano, JR, M Parham, and LH Hall. 2004. The Text Mining Hand. https://doi.org/10.1017/CBO978051154691 4.

Wahyu, Vipy. 2014. "Analisis Penerapan Algoritma Naive Bayes Dalam Pengklasifikasian Konten Berita Bahasa Indonesia." Universitas Dian Nuswantoro Semarang, no. 5: 5–6.

Wibowo, Ari Putra, and Eny Jumiati. 2017. "Sentiment Analysis Masyarakat Pekalongan Terhadap Pembangunan Jalan Tol Pemalang-Batang Di Media Sosial," no. 0285.

Wijaya, Akhmad Pandhu, and Heru Agus Santoso. 2016. "Naive Bayes Classification Pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government" 1 (1): 48–55.

Witten, Ian H. 2004. "Text Mining." The Practical Handbook of Internet Computing, 14-1-14–22. https://doi.org/10.1201/9780203507223.

Wong, AH, and L Abednego. 2015. "Ngelompokan Dokumen Otomatis Dengan Menggunakan T FIDf Classifier, Naive Bayes Classifier Dan KNN."

Wongso, Rini, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli, and Rudy. 2017. "News Article Text Classification in Indonesian Language." Procedia Computer Science 116: 137–43. https://doi.org/10.1016/j.procs.2017.10.039.

Yang, Christopher C., Hsinchun Chen, and Kay Hong. 2003. "Visualization of Large Category Map for Internet Browsing." Decision Support Systems 35 (1): 89–102. https://doi.org/10.1016/S0167-9236(02)00101-X.

Yang, Yiming, and Xin Liu. 1999. "A Re-Examination of Text Categorization Methods."

Zakzouk, Tarik S., and Hassan I. Mathkour. 2012. "Comparing Text Classifiers for Sports News." Procedia Technology 1: 474–80. https://doi.org/10.1016/j.protcy.2012.02.104.