# **BAB II**

**TINJAUAN PUSTAKA**

## Pemberitaan COVID-19

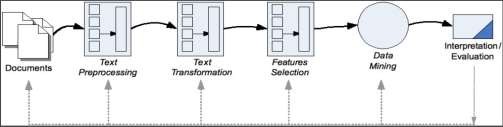
*Corona Virus Disease 2019* (COVID-19) adalah penyakit yang menginfeksi sistem pernafasan, disebabkan oleh jenis *coronavirus* baru yaitu SARS-CoV-2 yang terpusat di kota Wuhan, Cina, pada akhir Desember 2019. Setelah itu, COVID-19 menular antarmanusia dengan sangat cepat dan menyebar ke puluhan negara termasuk Indonesia, hanya dalam beberapa bulan.

Setelah masuknya COVID-19 di Indonesia, penyakit ini memberikan dampak ke bidang ekonomi, pendidikan, politik, pariwisata dan lainnya. tentunya media masa termasuk portal berita online dibanjiri dengan artikel-artikel berita tentang COVID-19.

## *Text Mining*

*Text Mining* adalah proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah besar sumber data tak terstruktur. Penambangan teks memiliki tujuan dan menggunakan proses yang sama dengan penambangan data, namun memiliki masukan yang berbeda. Masukan untuk penambangan teks adalah data yang tidak (atau kurang) terstruktur, seperti dokumen Word, PDF, kutipan teks, dll., sedangkan masukan untuk penambangan data adalah data yang terstruktur (Feldman & Sanger, 2007). Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur ini dengan menggunakan teknik dan alat yang sama dengan penambangan data.

Tahapan *Text Minng*



Gambar 2.1 Proses *Text Mining*

Area penerapan penambangan teks yang paling populer adalah:

1. Ekstraksi informasi (information extraction): Identifikasi frasa kunci dan keterkaitan di dalam teks dengan melihat urutan tertentu melalui pencocokan pola.
2. Pelacakan topik (topic tracking): Penentuan dokumen lain yang menarik seorang pengguna berdasarkan profil dan dokumen yang dilihat pengguna tersebut.
3. Perangkuman (summarization): Pembuatan rangkuman dokumen untuk mengefisienkan proses membaca.
4. Kategorisasi (categorization): Penentuan tema utama suatu teks dan pengelompokan teks berdasarkan tema tersebut ke dalam kategori yang telah ditentukan.
5. Penggugusan (clustering): Pengelompokan dokumen yang serupa tanpa penentuan kategori sebelumnya (berbeda dengan kategorisasi di atas).
6. Penautan konsep (concept linking): Penautan dokumen terkait dengan identifikasi konsep yang dimiliki bersama sehingga membantu pengguna untuk menemukan informasi yang mungkin tidak akan ditemukan dengan hanya menggunakan metode pencarian tradisional.
7. Penjawaban pertanyaan (question answering): Pemberian jawaban terbaik terhadap suatu pertanyaan dengan pencocokan pola berdasarkan pengetahuan(Marfian, 2015).

## Klasifikasi Teks

Klasifikasi merupakan tugas data mining yang memetakan data ke dalam kelompok-kelompok kelas. Teknik klasifiasi melakukan pengklasifikasian item data ke label kelas yang telah ditetapkan, membangun model klasifikasi dari kumpulan data yang dimasukkan, membangun model yang digunakan untuk memprediksi tren data masa depan. Algoritma yang umum digunakan meliputi K-Nearest Neighbor, Naïve Bayes Classification, Pohon Keputusan (Decision Tree), Jaringan Saraf (Neural Network), dan Suport Vector Machines (Sahu, Shrma, & Gondhalakar, 2011).

## *Preprocesing*

Tahap Praproses dokumen atau preprocessing merupakan tahapan utama yang dilakukan untuk menyiapkan dokumen menjadi lebih berkualitas agar menjadi data yang dapat diolah lebih lanjut. Tahapan praproses yang sering dilakukan pada dokumen yaitu cleansing, case folding, tokenization, stopword removal/filtering dan stemming.

## *Cleansing*

*Cleansing* merupakan proses pembersihan teks dari karakter yang tidak diperlukan yang menyebabkan adanya noise. Untuk karakter yang dihapuskan yaitu selain huruf ‘a’ sampai ‘z’ seperti symbol, tanda baca, dan juga url yang tidak memiliki arti sehingga dapat menggangu proses klasifikasi. Untuk menghilangkan karakter tersebut akan diubah menjadi karakter spasi.

## *Tokenizing*

*Tokenizing* adalah proses untuk membagi dokumen teks yang dapat berupa kalimat atau paragraf menjadi token-token atau bagian-bagian tertentu (Manning, Raghavan, & Schutze, 2009). Tokenizing menggunakan karakter spasi untuk memisahkan tiap kata dalam sebuah kalimat

## *Stopword Removal*

Biasanya, kata-kata yang sering muncul di setiap kategori dokumen merupakan katakata yang tak bermakna dalam pengkategorian dokumen (Manning, Raghavan, & Schutze, 2009). Stopword adalah kata yang biasanya muncul dalam kumpulan teks dan tidak mempunyai makna ataupun informasi yang bisa membantu dalam proses klasifikasi. Sehingga dengan menghapus stopword diharapkan yang tersisa dari suatu dokumen adalah kata-kata yang penting (Ramadhani, 2018)

## *Stemming*

*Stemming* adalah proses untuk menghilangkan imbuhan suatu kata sehingga menjadi sebuah kata dasar. Sebuah kata memiliki variasi imbuhan kata yang beragam. Variasi imbuhan dapat berupa awalan (prefiks), akhiran (sufiks), sisipan (infiks), dan kombinasi awalan dan akhiran (konfiks). Contoh proses pemotongan imbuhan dalam Bahasa Indonesia adalah kata “mengayuh” memiliki bentuk dasar “kayuh”, kata “memukul” memiliki bentuk dasar “pukul”, dan kata “beraksi” memiliki bentuk dasar “aksi” (Ramadhani, 2018). Penggunaan stemming akan mengurangi jumlah kata atau fitur, karena beberapa kata yang dipotong berkemungknan menghasilkan kata dasar yang sama (Garnes, 2009).

## Klasifikasi Teks

Klasifikasi merupakan tugas data mining yang memetakan data ke dalam kelompok-kelompok kelas. Teknik klasifiasi melakukan pengklasifikasian item data ke label kelas yang telah ditetapkan, membangun model klasifikasi dari kumpulan data yang dimasukkan, membangun model yang digunakan untuk memprediksi tren data masa depan. Algoritma yang umum digunakan meliputi K-Nearest Neighbor, Naïve Bayes Classification, Pohon Keputusan (Decision Tree), Jaringan Saraf (Neural Network), dan Suport Vector Machines (Sahu, Shrma, & Gondhalakar, 2011).

## Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah cara pemberian bobot hubungan suatu kata (term) terhadap dokumen. Untuk dokumen tunggal tiap kalimat dianggap sebagai dokumen (Mulyana, Ramadona, & Herfina, 2012). Metode ini menggabungkan dua konsep untuk perhitungan bobot, yaitu term frequency (TF) merupakan frekuensi kemunculan term (t) pada kalimat (d) dan document frequency (DF) adalah banyaknya kalimat dimana suatu term (t) muncul. Bobot suatu term semakin besar jika sering muncul pada suatu dokumen, dan semakin kecil jika muncul dalam banyak dokumen. Skema persamaan TF-IDF ditunjukkan oleh persamaan 2.1 (Naf'an, Burhanuddin, & Riyani, 2019).

𝑡𝑓𝑖𝑑𝑓(𝑤) = 𝑡𝑓 × (log ( 𝑁 𝑑𝑓) + 1) (2.1)

Keterangan:

Tf : Term frequency (jumlah kemunculan suatu term t dalam suatu dokumen d)

df : Document frequency (jumlah dokumen yang mengandung suatu term t)

N : Jumlah dokumen

tfidf(w): bobot term t terhadap dokumen d

Pada penelitian ini digunakan TFIDF pada library Scikit-learn. TFIDF Scikit-learn dilakukan dengan memberikan nilai False pada smooth IDF dan norm. Pemberian nilai False pada smooth IDF bertujuan untuk melakukan penyelesaian perhitungan TFIDF menggunakan persamaan 2.1. Pemberian nilai False pada norm bertujuan untuk mencegah Scikit-learn melakukan normalisasi pada hasil perhitungan TFIDF. Perhitungan log pada TFIDF Scikit-learn dilakukan dengan base 2,72 (logaritma natural)

## *Naive Bayes*

Naïve Bayes Classifier adalah merupakan algoritma klasifikasi machine learning (teknik pembelajaran) yang berdasarkan probabilitas sederhana. Algoritma Naïve Bayes Classifier juga berasal dari teorema Bayesian dengan asumsi independent yang cukup tinggi. Naïve Bayes Classifier berasumsi bahwa keberadaan variable satu tidak berkaitan sama sekali dengan keberadaan variable lainnya. Walaupun dalam penerapannya. Naïve Bayes Classifier merupakan suatu algoritma klasifikasi yang sederhana tapi performa dan tingkat akurasinya tergolong tinggi. Kelebihan Naïve Bayes Classifier adalah mudah untuk diimplementasikan karena modelnya yang sederhana. Selain itu metode Naïve Bayes Classifier juga tidak membutuhkan jumlah data yang terlalu besar untuk menentukan estimasi dari parameter yang akan dilakukan dalam proses klasifikasi. Pada metode ini, Teorema Bayes memiliki bentuk umum sebagai berikut :

P( H | X ) = (2.2)

Keterangan :

X = adalah data sampel dengan label yang belum diketahui

H = adalah hipotesa yang mengasumsikan X merupakan data dengan label.

P(H) = adalah peluang atau probablitas awal dari hipotesis H.

P(X) = adalah peluang dari data sampel.

P(X|H) = adalah peluang dari data sampel X.

P(H|X) = adalah peluang bahwa hipotesa benar untuk setiap data sampel X.

Dalam algoritma NBC setiap dokumen direpresentasikan dengan pasangan atribut “a1, a2, a3,..., an” dimana a1 adalah kata pertama, a2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori berita.

Pada saat klasifikasi algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (VMAP) (Dio Ariadi, 2015).

## Confusion Matrix

Pengukuran ketepatan klasifikasi sangat perlu dilakukan karena metode ini digunakan untuk menghitung kinerja akurasi hasil klasifikasi yang digunakan pada penelitian ini. Sistem yang dilakukan kinerja klasifikasi diharapkan mampu menghasilkan akurasi yang sangat tinggi, tetapi tidak dapat dipungkiri bahwa tidak mungkin sebuah kinerja sebuah sistem mencapai 100% benar. Maka mutlak harus diukur kinerja suatu sistem klasifikasinya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matrik konfusi (confusion matrix).Rumus dalam penghitungan akurasi adalah sebagai berikut:

Akurasi = (2.3)

Keterangan:

TP : keseluruhan data yang benar yang terklasifikasi dalam kelas

TN : keseluruhan data yang benar yang terklasifikasi sebaliknya dalam kelas

FP : keseluruhan data dianggap sistem benar berada dalam kelas namun seharusnya data tersebut berada dalam kelas sebaliknya

FN : keseluruhan data yang dianggap sistem berada dalam kelas sebaliknya namun seharusnya data tersebut berada dalam kelas yang benar..

## Tinjauan Literatur

Berbagai penelitian yang telah dilakukan dalam bidang *text mining* seperti *analysis sentiment,* analisis berita hoaks, klasifikasi teks berita **s**udah banyak dilakukan. Dalam penelitian mengenai klasifikasi konten berita menggunakan algoritma *naïve bayes* dibutuhkan referensi atau penelitian sebelumnya yang terkait untuk mendapat hasil yang maksimal. Berikut merupakan penelitian terkait yang sudah pernah dilakukan.

1. Penelitian yang dilakukan oleh (Bambang Kurniawan, 2012) melakukan penelitian mengenai **Klasifikasi Konten Berita Menggunakan Metode *Text Mining* Dengan Algoritma *Naïve Bayes***. Penelitian ini dilakukan karena pengunggah berita dalam membagi berita ke dalam kategori masih menggunakan cara manual, cara seperti ini sangat merepotkan bagi pengunggah berita apabila jumlah berita yang ingin diunggah berjumlah banyak. Berdasarkan masalah tersebut peneliti melakukan penelitian dan membuat sistem berbasis web yang dapat mengklasifikasikan berita-berita tersebut secara otomatis. Pada penelitian ini Data yang digunakan adalah data berita yang diperoleh dari beberapa media online. Data berita yang digunakan berjumlah 400 data berita lalu dibagi menjadi 4 konten berita yaitu berita ekonomi, berita politik, berita entertainment dan berita olahraga, masing-masing kategori tersebut berjumlah 100 data berita. Dari 100 berita itu 90 berita dijadikan sebagai data training dan 10 data berita dijadikan sebagai data testing. Kesimpulan dari penelitian ini yaitu aplikasi yang telah dibuat mampu melakukan proses klasifikasi data berita secara otomatis namun proses klasifikasi dapat lebih akurat jika data latih yang digunakan dalam pembelajaran berjumlah banyak.
2. Dalam penelitian oleh (Nurhadi, 2016) melakukan penelitian mengenai **Implementasi Algoritma *Naïve Bayes Classifier* Berbasis *Particle Swarm Optimization*** untuk mengkategorikan konten berita digital Bahasa Indonesia. Dalam penelitian ini, penulis melakukan pengujian model menggunakan teknik 10 cross validation, di mana proses ini membagi data secara acak ke dalam 10 bagian. Pengujian dimulai dengan pembentukan model dengan data pada bagian pertama. Model yang terbentuk akan diujikan pada 9 bagian data sisanya. Setelah itu proses akurasi dihitung dengan melihat seberapa banyak data yang sudah terklasifikasi dengan benar. Hasil dari pengujian model yang dilakukan adalah mengklasifikasikan berita gosip, berita travel, dan berita kuliner dari suatu konten berita dengan Naïve Bayes Classifier berbasis Particle Swarm Optimization untuk menentukan nilai accuracy. Pengukuran akurasi menggunakan confusion matrix terbukti bahwa hasil pengujian algoritma Naïve Bayes Classifier berbasis Particle Swarm Optimization memiliki nilai akurasi yang lebih tinggi. Dengan nilai akurasi sebesar 94.17% yang didapat saat proses 5 fold cross validation.
3. Penelitian mengenai **Analisis Penerapan Algoritma *Naive Bayes* Dalam Pengklasifikasian Konten Berita Bahasa Indonesia** yang dilakukan oleh (Vipy Wahyu Perdana, 2014). Penelitian ini dilakukan dengan tujuan untuk membantu dan mempermudah kinerja dari editor berita. Hasil yang didapat dari penelitian ini yaitu Tf-idf dan *Naive Bayes* mampu melakukan proses klasifikasi 6 kategori berita dengan baik. Dari pengujian klasifikasi kategori berita ini penggunaan Tf-Idf dan Naive Bayes dapat digunakan untuk proses klasifikasi berita dengan nilai akurasi sebesar 89.22%.
4. Penelitian yang dilakukan (Dio Ariadi, 2015) melakukan penelitian mengenai **Klasifikasi Berita Indonesia Menggunakan Metode *Naive Bayesian Classification* Dan *Support Vector Machine* Dengan *Confix Stripping Stemmer***. Latar belakang dilakukannya penelitian ini yaitu pengkategorian berita yang masih menggunakan tenaga manusia atau manual, cara seperti ini tentu akan menyulitkan pihak editor berita untuk mengkategorikan, terutama untuk berita yang tidak terlalu berbeda secara jelas. Dalam penelitian ini dilakukan dengan menggunakan dua metode, metode pertama adalah yang umumnya dipakai yaitu metode *Naïve bayes* dan metode kedua adalah metode SVM. Kedua metode tersebut akan dibandingkan, mana metode yang menghasilkan tingkat klasifikasi paling besar. Kesimpulan yang didapat dari penelitian ini yaitu Metode Naive Bayes Classifier dapat melakukan klasifikasi berita Indonesia cukup baik. Hasil yang didapatkan pada saat data testing pada masing-masing pengukuran performa akurasi, precision, recall, dan FMeasure sebesar 82,2%; 83,9%; 82,2%; dan 82,4%. 2. Metode Support Vector Machine antara kernel RBF dan kernel linier pada word vector 10000sama baiknya dalam melakukan klasifikasi berita Indonesia. Menggunakan data testing didapatkan untuk tiap pengukuran performa akurasi, precision, recall, dan FMeasure adalah 88,1%, 89,1%, 88,1%, dan 88,3%.

**Tabel 2.1** Perbandingan Penelitian Terkait

|  |  |  |  |
| --- | --- | --- | --- |
| **Judul Penelitian** | **Peneliti dan Tahun Penelitian** | **Metode** | **Hasil Penelitian** |
| Klasifikasi Konten Berita Dengan Metode Text Mining | (Bambang Kurniawan, Syahril Effendi, Opim Salim Sitompul, 2012) | Text Mining | Dari hasil dari penelitian ini dapat disimpulkan bahwa Aplikasi yang dibuat mampu melakukan proses klasifikasi data berita secara otomatis dan proses klasifikasi dapat lebih akurat jika data latih yang digunakan dalam pembelajaran ditambah banyak. |
| Analisis Penerapan Algoritma Naive Bayes dalam Pengklasifikasian Konten Berita Bahasa Indonesia | (Vipy Wahyu Perdana, Heru Agus Santoso, 2014) | Naïve Bayes | Hasil dari penelitian ini yaitu Tf-idf dan algortima Naive Bayes mampu melakukan proses klasifikasi 6 kategori berita ini dengan baik. Rata – rata nilai recall yang dihasilkan sebesar 89.23%, dan rata – rata nilai precision sebesar 89.44%. Serta nilai akurasi dari klasifikasi ini mencapai 89.22%. |
| Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer | (Dio Ariadi, Kartika Fithriasari, 2015) | Naïve Bayes, Support Vector Machine | Hasil dari pengujian dengan Metode Naive Bayes Classifier dapat melakukan klasifikasi berita Indonesia cukup baik. Hasil yang didapatkan pada saat data testing pada masing-masing pengukuran performa akurasi, precision, recall, dan FMeasure sebesar 82,2%; 83,9%; 82,2%; dan 82,4%. 2. Metode Support Vector Machine antara kernel RBF dan kernel linier pada word vector 10000 sama baiknya dalam melakukan klasifikasi berita Indonesia. Menggunakan data testing didapatkan untuk tiap pengukuran performa akurasi, precision, recall, dan FMeasure adalah 88,1%, 89,1%, 88,1%, dan 88,3%. |
| Implementasi Algoritma Naïve Bayes Classifier Berbasis Particle Swarm Optimization (PSO) Untuk Klasifikasi Konten Berita Digital Bahasa Indonesia | (Acmad Nurhadi, 2016) | Naïve Bayes | Hasil dari pengujian yang dilakukan untuk mengklasifikasikan berita dengan kategori gosip, travel, dan kuliner dari suatu konten berita dengan algoritma Naïve Bayes Classifier berbasis Particle Swarm Optimization diperoleh hasil akurasi sebesar 94.17% yang didapat saat proses 5 fold cross validation. |

.