

# BAB 1 PENDAHULUAN

## 1.1 Latar Belakang

Covid-19 menjadi permasalahan internasional beberapa tahun kebelakang ini. Virus yang menyerang pernapasan ini mulai berdampak bagi kehidupan masyarakat dari akhir tahun 2019 (dr. Iqbal, 2022). Hal ini membuat masyarakat mencari banyak informasi tentang virus ini, dan membuat covid-19 menjadi ramai diperbincangkan di sosial media, salah satunya *twitter*.

*Twitter* mempunyai fitur *Trending Topics*. Fitur ini digunakan untuk mengetahui informasi terbaru di seluruh dunia secara *real time*. Dengan jumlah pengguna harian mencapai 126 juta per hari (Yudha, 2019), *twitter* menjadi salah satu sumber informasi atau konten tercepat. Dibarengi dengan penggunaan *hashtag* yang digunakan untuk membicarakan berita atau informasi yang sedang viral, dengan cara ini mendapatkan informasi tentang Covid-19 menjadi lebih mudah. Terlebih lagi saat ini pemerintah sedang mengejar target untuk mem-vaksin rakyat Indonesia. Dengan menduduki peringkat 6 penggunaan *twitter* terbanyak di Dunia tahun 2021 (Dinda, 2021), masyarakat Indonesia akan memanfaatkan fitur ini.

Fitur *trending topics* ini menampung semua informasi yang paling sering dibicarakan, dalam hal ini covid-19. Ini menyebabkan pengguna perlu menyisihkan waktu untuk memilih informasi yang didapatkan, dan diinginkan. Informasi yang tidak diperlukan secara terus menerus muncul, dan tidak memiliki keterkaitan dengan informasi yang dicari dapat dikategorikan sebagai *spam*.

Untuk mempermudah pengguna maka dari itu dibutuhkan suatu sistem yang dapat mengklasifikasi tweet ke dalam 2 kategori yaitu *spam* dan *non spam*. Sistem tersebut dapat dibangun dengan berbagai metode dan algoritma. Salah satunya menggunakan algoritma *Naïve Bayes*. Algoritma *Naïve Bayes* adalah algoritma yang dapat mengolah data dalam jumlah besar dengan akurasi yang cukup tinggi. *Naive Bayes* adalah salah satu metode yang populer yang dapat digunakan untuk mengklasifikasikan dokumen secara probabilistic. Metode *Naïve Bayes* adalah jenis dari Teori Bayesien dimana kondisi atau kelas yang ada adalah independen dan tidak terikat satu sama lain (Faisal, Aryo, & Rizal, 2018). Oleh karena itu penulis, akan mengembangkan suatu sistem klasifikasi spam tweet covid-19 menggunakan algoritma *Naïve Bayes*.

## 1.2 Perumusan Masalah

Dari uraian latar belakang diatas, maka yang menjadi permasalahan utama adalah sebagai berikut:

1. Bagaimana penerapan *naïve bayes* bisa mengklasifikasi *tweet* ke dalam 2 kategori *spam* dan *non spam* berdasarkan *hashtag* dan *tweet* yang bisa diklasifikasi yang berbahasa Indonesia.
2. Bagaimana agar dapat meningkatkan perhitungan akurasi dalam sebuah sistem klasifikasi *spam* dan *non spam* pada *twitter*.

## 1.3 Tujuan Penelitian

Berikut merupakan tujuan dari sistem ini sebagai tugas akhir, yaitu:

1. Mengimpelentasikan cara kerja algoritma Naïve bayes untuk membangun sistem klasifikasi tweet menjadi 2 kategori spam jika isi tweet tidak sesuai dengan hastag dan non spam jika isi tweet sesuai dengan hastag dalam hal ini #covid19 dengan Bahasa Indonesia.
2. Mengetahui kinerja Naïve Bayes untuk melakukan klasifikasi *spam*.

## 1.4 Batasan Masalah

Agar pengerjaan tugas akhir ini dapat terarah, terimplementasi dengan baik, dan mendapatkan hasil yang lebih spesifik dan jelas, maka sistem yang dirancang akan dibatasi dengan ruang lingkup sebagai berikut:

1. Sistem ini dirancang untuk mengklasifikasi *tweet* dengan bahasa Indonesia.
2. Tidak mengidentifikasi gambar, video, link, dan Simbol.
3. Menggunakan data set yang berasal dari *tweet* sosial media twitter.
4. Tidak menggunakan operator atau proses *stemming*.
5. Mengklasifikasi *tweet* berdasarkan *hashtag*.

## 1.5 State of The Art

**Tabel 1. 1** *State of The Art*

| Jurnal   | Kesimpulan   | Persamaan   | Perbedaan   |
|--|--|---|---|
| Rancang Bangun Aplikasi Deteksi <i>Spam</i> <i>Twitter</i> | Dalam penelitian ini, penulis merancang sebuah aplikasi android untuk mendeteksi <i>spam</i> | <ul style="list-style-type: none"><li>• Menggunakan metode yang sama yaitu <i>Naïve Bayes</i></li></ul> | <ul style="list-style-type: none"><li>• Melakukan perbandingan dengan metode lain</li></ul> |

|   |   |   |   |
|---|---|---|---|
| <p>menggunakan Metode <i>Naïve Bayes</i> dan KNN pada Perangkat Bergerak Android</p> <p>(Faisal Aji Prayoga, Aryo Pinandito, Rizal Setya Perdana, 2018)</p>                   | <p><i>twitter</i> menggunakan metode <i>Naïve Bayes</i> dan KNN. Melalui indentifikasi dengan <i>data set</i> 10 hingga 100 menggunakan dua metode tersebut dapat mendeteksi spam dengan akurasi masing-masing 82% dan 71%.</p>   |   | <ul style="list-style-type: none"> <li>• Merancang aplikasi android</li> </ul>                          |
| <p>Penerapan Metode <i>Naïve Bayes</i> Untuk Klasifikasi SMS <i>Spam</i> Menggunakan <i>Java Programming</i></p> <p>(Eko Ardian Pranata, Subari, Go Frendi Gunawan, 2019)</p> | <p>Dalam penelitian ini, penulis ingin mengklasifikasi <i>spam</i> pada SMS dengan menggunakan <i>Java Programming</i> dengan melakukan perhitungan tingkat kemiripan isi dari SMS yang belum memiliki kategori dalam hal ini SMS <i>spam</i> dan bukan <i>spam</i>. Yang menjadi <i>data training</i> adalah SMS yang telah dimasukan ke dalam <i>database</i>. Dan mendapatkan hasil pengujiann akurasi</p> | <ul style="list-style-type: none"> <li>• Menggunakan metode yang sama yaitu <i>Naïve Bayes</i></li> </ul> | <ul style="list-style-type: none"> <li>• Melakukan perhitungan tingkat kemiripan antar teks.</li> </ul> |

|   |  |   |  |
|---|--|---|--|
|   | sebesar 62% dari 200 data latih dan 200 data uji.pak   |   |  |
| <p>Deteksi <i>Spam</i> pada <i>Twitter</i></p> <p>Menggunakan Algoritme <i>Naïve Bayes</i></p> <p>(Anindita Wahyuningtyas, Imas Sukaesih Sitanggang, Husnul Khotimah, 2020)</p> | <p>Dalam penelitian ini, penulis menggunakan algoritma <i>Naïve Bayes</i> untuk mengklasifikasi <i>spam</i> pada <i>twitter</i> dengan hasil yang cukup baik dan akurasi yang tinggi yaitu 95.57%.</p> <p>Dalam penelitian ini penulis mengabaikan duplikat <i>tweet</i> yang berarti jika ada <i>tweet</i> yang berisi sama, akan dianggap sebagai <i>tweet</i> yang berbeda.</p> | <ul style="list-style-type: none"> <li>• Algoritma yang digunakan sama yaitu <i>Naïve Bayes</i>.</li> <li>• Data yang didapat melalui sumber yang sama yaitu <i>twitter</i> API.</li> </ul> | <ul style="list-style-type: none"> <li>• Data yang diambil telah ditentukan yaitu 25 akun <i>spammer</i> dan 25 akun bukan <i>spammer</i>.</li> <li>• Tidak membutuhkan <i>Hashtag</i> untuk mendeteksinya.</li> <li>• Mendeteksi <i>tweet</i> berbahasa Inggris.</li> </ul> |
| <p>Klasifikasi Spam Pada Twitter</p> <p>Menggunakan Metode <i>Improved K-Nearest Neighbor</i></p>   | <p>Dalam penelitian ini, penulis menggunakan metode <i>Improved K-Nearest Neighbor</i>.</p> <p>Pada penelitian ini penulis menggunakan data latih sebanyak 500 dokumen dan mendapatkan akurasi</p>   | <ul style="list-style-type: none"> <li>• Menggunakan tahapan <i>cleansing, case folding</i>.</li> </ul>   | <ul style="list-style-type: none"> <li>• Metode yang digunakan menggunakan metode <i>Improve K-Nearest Neighbor</i></li> <li>• Menggunakan perhitungan</li> </ul>  |

|  |   |  |   |
|--|---|--|---|
| (Dea Zakia Nathania, Indriati, Fitra Abdurrachma Bachtiar, 2018)   | untuk metode <i>Improved K-Nearest Neighbor</i> sebesar 92% dibandingkan menggunakan metode <i>K-Nearest Neighbor</i> hanya sebesar 88%.  |  | <i>Cosine Similarity</i> .  |
| Mendeteksi <i>Spammers</i> di Twitter dengan <i>SVM Classifier</i> | Dalam penelitian ini, penulis menggunakan metode <i>SVM Classifier</i> , dengan menggunakan Bahasa pemrograman PHP, serta menggunakan metode <i>percentage split</i> dan menghasilkan akurasi yang baik sebesar 96.67% dengan rasio 90:10 untuk <i>data training : data testing</i> . | <ul style="list-style-type: none"> <li>• <i>Hashtag</i> merupakan salah satu cara untuk mendeteksi <i>spam</i> jika <i>hashtag</i> dan isinya tidak berhubungan dianggap <i>spam</i>.</li> </ul> | <ul style="list-style-type: none"> <li>• Metode yang digunakan adalah metode <i>SVM Classifier</i>.</li> <li>• Menggunakan bahasa pemrograman PHP.</li> </ul> |

## 1.6 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran secara umum mengenai permasalahan dan pemecahannya. Penyusunan ini diuraikan dalam beberapa permasalahan yang terbagi menjadi beberapa bab sebagai berikut:

### **Bab 1           Pendahuluan**

Bab ini memuat pendahuluan penelitian yang terdiri dari latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian,

ruang lingkup penelitian, *state of the art*, dan sistematika penulisan penelitian.

**Bab 2      Landasan Teori**

Bab ini memuat landasan teori penelitian yang terdiri dari teori dasar mengenai *Machine Learning*, *twitter*, *spam*, algoritma *Naïve Bayes*, dan TF-IDF

**Bab 3      Metode**

Bab ini memuat metode yang digunakan untuk membuat sistem klasifikasi *spam tweet* pada *twitter*.

**Bab 4      Hasil dan Pembahasan**

Bab ini membahas hasil implementasi algoritma *Naïve Bayes* dalam mendeteksi dan mengklasifikasi *spam* maupun *non spam tweet*.

**Bab 5      Kesimpulan dan Saran**

Bab ini berisi kesimpulan dan saran dari pembahasan implementasi sistem klasifikasi *spam tweet* pada *twitter* dengan metode *Naïve Bayes*.