# DETEKSI UJARAN ANCAMAN BERBASIS WEBSITE PADA POSTINGAN MEDIA SOSIAL TWITTER MENGGUNAKAN METODE NAIVE BAYES

## *WEBSITE BASED DETECTION OF THREATS IN SOCIAL MEDIA TWITTER USING NAIVE BAYES METHOD*

**Andhika Rafi R[1], Muhammad Nasrun[2], Ratna Astuti N[3]**
Prodi S1 Teknik Komputer, Fakultas Teknik Elektro, Universitas Telkom
[1]**andhikarafir@student.telkomuniversity.ac.id**, [2]**muhammadnasrun@telkomuniversity.co.id**,
[3]**ratnaan@telkomuniversity.ac.id**

**Abstrak**

**Di era teknologi zaman sekarang media sosial sangat penting bagi kehidupan manusia. Media sosial berisi informasi yang digunakan untuk berkomunikasi, iklan, pendidikan, acara, dan dibutuhkan oleh kegiatan manusia untuk bersosialisasi. Di jaman yang berkembang seperti sekarang ini media sosial banyak di salah gunakan oleh oknum-oknum yang tidak bertanggung jawab, salah satunya menyebarkan sebuah pesan ancaman.Penulis melihat media sosial twitter sangat banyak postingan yang mengandung pesan ancaman yang ditujukan untuk pemerintahan, kelompok, bahkan individu. Sarana, platform, atau aplikasi sangat dibutuhkan untuk dapat mengidentifikasi sebuah kata atau kalimat yang mengandung sebuah pesan negatif khususnya ancaman. Pada penelitian ini dibuat sebuah sistem yang akan membantu pemerintah bahkan institusi terkait berupa aplikasi website yang akan mendeteksi ujaran ancaman pada postingan media sosial twitter menggunakan metode *machine learning* berupa *naive bayes*. Hasil dari penelitian ini menunjukkan bahwa sistem pendeteksi ujaran ancaman pada twitter yang di buat mendapatkan akurasi sebesar 66%, *precision* 64%, *recall* 63%, dan F1 *score* sebesar 63%.**

**Kata kunci : ancaman, twitter, *Naive Bayes***

***Abstract***

***In today's technological era, social media is very important for human life. Social media contains information used for communication, advertising, education, events, and is needed by human activities to socialize. In this developing era, social media is often misused by irresponsible individuals, one of which is spreading a threatening message.The author sees in social media twitter there are many posts containing threatening messages aimed at governments, groups, and even individuals. Tools, platforms or applications are needed to be able to identify a word or sentence that contains a negative message, especially a threat. In this study, a system was created that would help the government and even related institutions in the form of a website application that would detect threat utterances in Twitter social media posts using machine learning methods in the form of Naive Bayes. The results of this study indicate that the threat speech detection system created on Twitter has an accuracy of 66%, 64% precision, 63% recall, and an F1 score of 63%.***

**Keywords:** ***threat, twitter, Naive Bayes***

## 1. Pendahuluan

Twitter merupakan salah satu media sosial populer di Indonesia. Situs resmi kominfo pada tahun 2013 menyatakan Indonesia menempati peringkat 5 pengguna Twitter terbesar di dunia setelah USA, Brazil, Jepang dan Inggris sebanyak 19,5 juta pengguna dari total 500 juta pengguna global[1]. Maraknya ancaman pada linimasa media sosial meningkatkan kewaspadaan kita mulai dari ancaman pembunuhan, ancaman pengerusakan, dan ancaman lainnya. Imbas dari sebuah ancaman adalah ketakutan, kegelisahan, dan kecemasan. Beberapa orang dalam media sosial banyak yang menyembunyikan identitasnya sendiri dan susah untuk ditelusuri pelakunya.

Kementerian Komunikasi dan Informatika mengingatkan agar berhati hati dalam melakukan aktivitas pada media sosial, mereka tidak bisa sembarangan menyebar kebencian, hoax, bahkan

suatu ancaman karena bisa terjerat Undang-Undang Informasi. Dalam UU ITE terdapat aturan dalam bermedia sosial, dengan begitu pengguna medsos tidak bisa seenaknya melakukan unggahan yang berisi ancaman atau intimidasi hingga persekusi. Maka dari itu dibutuhkan sebuah aplikasi atau platform yang bisa membatu pemerintah atau pihak berwenang mencari kata atau kalimat yang berisikan sebuah pesan ancaman yang pada umumnya banyak di media sosial diantaranya di Twitter.

Pada penelitian tugas akhir ini penulis membuat aplikasi website yang fokus pada perancangan untuk mendeteksi ujaran ancaman pada postingan media sosial twitter menggunakan metode *Naive Bayes* berbasis website. Pada perancangan ini penulis menggunakan metode dari *machine learning* yaitu *Naive Bayes*. *Naive Bayes* sendiri dipilih karena memberikan performansi cukup baik untuk banyak kasus modern dengan data yang besar. Dengan adanya sistem yang akan dibuat ini, ujaran ancaman di media sosial Twitter dapat di deteksi dan mempermudah lembaga terkait untuk menangani kasus maraknya ujaran ancaman di media sosial twitter.

## 2. Landasan Teori

### 2.1 *Multinomial Naive Bayes*

*Naive Bayes* adalah pendekatan probabilitik murni untuk klasifikasi, dimana idenya adalah menggunakan probabilitas, seberapa sering sebuah kata muncul pada kelas untuk menentukan probabilitas kalimat yang termasuk dalam kelas tertentu[9].

$$P(doc|Vj)\prod_{i=1}^{length(doc)} P(ai = Wk|Vj) \quad (1)$$

$$VNB = \underset{Vj \in V}{\operatorname{argmax}} P(Vj)\prod_{w \in words} P(W|Vj) \quad (2)$$

$$P(Wk|0/1) = \frac{nk+1}{n0/1+|Vocabulary|} \quad (3)$$

### 2.2 Twitter

Twitter merupakan salah satu media sosial populer di Indonesia. Situs resmi kominfo pada tahun 2013 menyatakan Indonesia menempati peringkat 5 pengguna Twitter terbesar di dunia setelah USA, Brazil, Jepang dan Inggris sebanyak 19,5 juta pengguna dari total 500 juta pengguna global[1]. Twitter merupakan media sosial yang penggunanya dapat menulis tweet dan berkomentar yang biasanya di sebut mention di kalangan penggunanya. Dari tweet dan mention tersebut, pengguna dapat memberikan pendapat dan keinginannya yang ingin mereka tulis di sebagia tweet atau di mention kepada orang lain. Dari tweet dan mention tersebut dapat memberikan celah adanya ujaran ancaman.

### 2.3 *PreProcessing*

Nomor urut tabel ditulis di bagian atas tabel yang dijelaskan, beikut ini contoh penulisan tabel: Tabel 1, Tabel 2(a). *Text preprocessing* merupakan proses mengubah bentuk data yang belum memliki struktur menjadi data yang terstruktur sesuai dengan kebutuhan, untuk proses *mining* yang lebih lanjut. Sebuah teks yang ada harus dipisahkan, hal ini dapat dilakukan dalam beberapa tingkatan yang berbeda. Pengubahan bentuk dapat berupa memecah paragraf menjadi kalimat dan kalimat akhirnya menjadi kata serta dapat menghilangkan angka, simbol atau karakter-karater lainya. Tahapan *preprocessing* berdasarkan meliputi: *case folding*, *tokenizing/parsing*, *filtering*, *stemming*.

*a. Case Folding*

*Case folding* adalah mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter. Contoh kata “KOMPUTER” akan menjadi “komputer”.

*b. Tokenizing*

*Tokenizing* yaitu pemotongan *string input* tiap kata yang menyusunnya. Contohnya : kalimat “Aku cinta Undiksha” dipotong menjadi kata : Aku | Cinta | Undiksha.

*c. Filtering*

*Filtering* adalah tahap mengambil kata - kata penting dari hasil *tokenizing*. Proses *filtering* dapat menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist*

(menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopword* adalah "yang", "dan", "di", "dari" dan lain – lain.

*d. Stemming*

*Stemming* merupakan suatu proses yang mentransformasikan kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (*root word*) dengan menggunakan aturan-aturan tertent). Pada tahapan ini menggunakan *stemming* untuk teks berbahasa Indonesia.

**2.4 Klasifikasi Pengujian**

Pengujian dilakukan dengan mengukur performansi dari model klasifikasi, diukur dari perbandingan antara data latih dengan dengan data uji. Untuk mengukur performansi suatu model terdapat 4 parameter yaitu *precision, recall, F1 score*, dan *accuracy*. Parameter ini didapatkan dengan membandingkan data uji dengan data latih dari hasil validasi dengan Balai Bahasa[6].

*Accuracy* merupakan tingkat kedekatan dari data latih dan data uji. Persamaan untuk menghitung *accuracy* yaitu:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

*Precision* merupakan tingkat ketepatan dari data yang diminta oleh pengguna dengan data yang dihasilkan oleh sistem. Persamaannya yaitu:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

*Recall* merupakan tingkat keberhasilan sistem dalam mengklasifikasikan. Persamaannya yaitu:

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

*F1-Score* merupakan evaluasi yang terdiri dari gabungan antar *precision* dan *recall.* Persamaannya yaitu:

$$F1\text{- }score = 2\ x\ \frac{precision\ x\ recall}{precision+recall} \quad (7)$$

Ket:

TP = Data yang diklasifikasikan sebagai ujaran ancaman

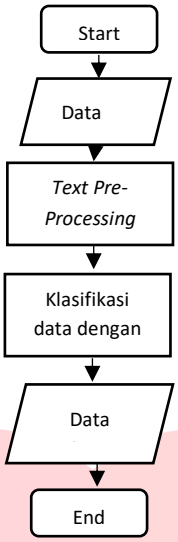
TN = Data yang bukan ujaran ancaman tetapi sistem mengklasifikasikan sebagai ujaran ancaman

FP = Data yang merupakan ujaran ancaman tetapi sistem mengklasifikasikan bukan ujaran ancaman

FN = Data yang diklasifikasikan bukan ujaran ancaman

## 3. Pembahasan

### 3.1. Perancangan Sistem




Gambar 1. Rancangan Sistem

Pada gambar 1 menunjukkan Rancangan Sistem. Pada gambar tersebut adalah gambaran dari sistem program yang di mulai dengan memasukkan kata pada kolom di sediakan, lalu program akan mulai melakukan pemrosesan *crawling* data dari twitter. Kalimat masuk ke proses *preprocessing* untuk selanjutnya di klasifikasikan dengan metode *naive bayes*, dari hasil memprosesan akan didapakan beberapa daftar tweet yang mengandung unsur ujaran ancaman sesuai kata yang dimasukkan.

### 3.2 Pengujian Data

Tabel 1. Rangkuman Pengujian Pasrtisi Data

| **Pengujian ke-** | **Data Uji (%)** | **Data Latih (%)** | ***Precision (%)*** | ***Recall (%)*** | ***F-1 Score (%)*** | ***Accuracy (%)*** |
|---|---|---|---|---|---|---|
| **1** | 10 | 90 | 64 | 63 | 63 | 66 |
| **2** | 20 | 80 | 61 | 60 | 61 | 64 |
| **3** | 30 | 70 | 62 | 61 | 61 | 64 |
| **4** | 40 | 60 | 60 | 60 | 60 | 63 |
| **5** | 50 | 50 | 55 | 57 | 57 | 61 |
| **6** | 50 | 40 | 56 | 55 | 54 | 60 |
| **7** | 70 | 30 | 56 | 55 | 55 | 60 |
| **8** | 80 | 20 | 55 | 54 | 54 | 59 |
| **9** | 90 | 10 | 49 | 49 | 47 | 56 |

Pada tabel 1 dapat dilihat bahwa pengujian pertama memperoleh tingkat *accuracy* tertinggi yaitu 66% yang dimana nantinya partisi tersebut akan dipakai untuk pengujian selanjutnya. Karena semakin banyak data latih dan semakin balance data maka semakin bagus akurasi sistem yang diperoleh

## 4. Kesimpulan Dan Saran

### 4.1 Kesimpulan

Berdasarkan hasil penelitian dan pengujian dan analisa yang telah dilakukan pada tugas akhir ini, maka dapat ditarik kesimpulan bahwa:

1. Sistem deteksi ujaran kebencian dalam Bahasa Indonesia Pada tweet dan mention di Twitter dengan menggunakan metode *Naive Bayes* berbasis website berhasil mengklasifikasikan tweet dan mention di twitter berupa kalimat ujaran ancaman dan bukan kalimat ujaran ancaman.
2. Berdasarkan pengujian partisi data, semakin banyak data dan balancenya jumlah data ancaman dengan yang bukan ancaman maka *accuracy* klasifikasi yang dihasilkan semakin bagus. Dalam hal ini partisi data yang terbaik pada proses pengujian yaitu 90% data latih data sedangkan 10% data uji.
3. Banyaknya data dan proses *preprocesing* mempengaruhi kinerja sistem dan *accuracy* yang dihasilkan. Dalam kasus ini nilai akurasi *final* pada proses pengujian yaitu sebesar 66%.Pada proses pengujian sistem diperoleh nilai rata-rata parameter *precision, recall,* dan *f-1 score* sebesar 63% dan *accuracy* sebesar 66%.

### 4.2 Saran

Berdasarkan hasil penelitian, pengujian dan analisa yang telah dilakukan pada tugas akhir ini, maka saran yang dapat diusulkan untuk penelitian lebih lanjut yaitu:

1. Implementasi dataset yang di gunakan harus balance.
2. Preprocessing dan klasifikasi yang di gunakan harus sesuai dengan data yang ada agar tidak mempengaruhi hasil

**Reference:**


[1] Selamatta Sembiring, "Kominfo: Pengguna Internet di Indonesia 63 Juta Orang," 2013. [Online]. Available: https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+internet+di+Indonesia+63+Juta+Orang/0/berita_satker [Accessed: 10-Jan-2021].

[2] Kuncahyo Setyo Nugroho, "*Confusion Matrix* untuk Evaluasi Model pada *Supervised Learning*" 2019. [Online]. Available: https://medium.com/@ksnugroho/confusion-matrix-untuk-evaluasi-model-pada-unsupervised-machine-learning-bc4b1ae9ae3f [Accessed: 10-Jan-2021].

[3] Rageeni Sah, "*Text Data Cleaning - tweets analysis*" 2018. [Online]. Available: https://www.kaggle.com/ragnisah/text-data-cleaning-tweets-analysis [Accessed: 10-Jan-2021].

[4] Suyanto, "*Machine Learning* Tingkat Dasar Dan Lanjut" 2018. Penerbit Informatika.

[5] Onno W. Purbo, "*Text Mining*" 2017. Penerbit Andi.

[6] Elvira Erizal, Budhi Irawan, and Casi Setianingsih, "*Hate Speech Detection in Indonesian Language on Instagram Comment Section Using Maximum Entropy Classification Method*," 2019 *Internasional Conference on Information and Communications Technology*.

[7] Ni Made Yeni Dwi Rahayu, "Rancangan Penerapan Metode Naive Bayes dalam Mendeteksi Hate Speech di Media Sosial," Prosiding Seminar Nasional Pendidikan Teknik Informatika, 2018.

[8] Muhammad Hakiem, Mochammad Ali Fauzi, and Indriati, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naive Bayes Berbasis N-Gram Dengan Seleksi Fitur *Information Gain*," Jurnal Pengembangan Teknologi Informatisi dan Ilmu Komputer, 2019.

[9] Erik Edward, “Comparing Methods of Text Categorization,” Examensarbete 15 hp, Juni 2018.
[10] Vivek Wisdom and Rajat Gupta, “An introduction to Twitter Data Analysis in Python,” 2016.
[11] R Kusumawati, A D 'Arofah, and P A Pramana, “Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services,” *Journal of Physics: Conference Series*, 2019.