JURNAL NASIONAL ILMU KOMPUTER

Jurnal Nasional Ilmu Komputer
e-ISSN: 2746-1343
Vol. 4, No.1,Februari 2023

Sentimen Pengguna Twitter Mengenai Isu Kebocoran Data Dengan Algoritma Naïve Bayes

Achmad Musyaffa Taufiqi[1], Aryo Nugroho[2].

Teknik Informatika, Fakultas Ilmu Komputer, Universitas Narotama
Email: amusfq@gmail.com[1], aryo.nugroho@narotama.ac.id[2]

Abstract

Sensitive data leakage occurs when it is unintentionally made available to or accessed by unauthorized individuals. Sentiment analysis is the process of analyzing text to gather different data from the web and social media in order to obtain opinions from internet users on a particular topic. Sentiment analysis is a branch of Natural Language Processing (NLP) that develops a system for extracting and identifying public opinion from text that is then transformed into structured data. One kind of technique for classifying data is Naïve Bayes. Naïve Bayes uses conditional probabilities to assign labels to instances. This study collects information from tweets sent by Twitter users on the subject of "Data Leakage" and then uses the Naïve Bayes method for sentiment analysis to determine the precision of GaussianNB.

Keywords: Sentiment Analisis, Nlp, Data Leakage, Naïve Bayes, Gaussiannb

1. Pendahuluan

1.1 Latar Belakang

Kebocoran data menjadi isu hangat yang ramai diperbincangkan oleh masyarakat di media sosial beberapa waktu lalu dimulai dari kebocoran data pengguna Indihome, PLN sampai data registrasi SIM Prabayar yang kemudian menyebabkan banyak peretas lain yang juga menyebarkan data yang mereka dapat dari beberapa instansi pemerintah.

Twitter merupakan media sosial yang popular dikalangan masyarakat selain Instagram dan Tiktok. Menurut data dari statista.com, Indonesia berada pada posisi kelima sebagai pengguna Twitter setelah India, Amerika, Jepang dan Brazil (S, n.d.).

Twitter sering digunakan masyarakat sebagai sarana untuk mengungkapkan perasaan terhadap suatu, baik melakui pujian maupun kritik (Zaskya et al., 2021). Karena itu Twitter dapat dimanfaatkan untuk menilai pandangan masyarakat terhadap suatu masalah atau objek. Emosi pada tweet pengguna dapat disebut sebagai analisa opini. Analisis sentimen juga digunakan untuk mengidentifikasi kecenderungan atau pendapat seseorang tentang suatu topik, apakah opini tentang topik tersebut positif atau negatif (Rozi et al., 2012).

2. Tinjauan Literatur

2.1 Data Mining

Data mining adalah proses mengolah kumpulan data dan menyajikan sebagai informasi. Metode yang digunakan *data mining* yaitu perhitungan seperti statistik, matematika, dan Kecerdasan buatan (*AI*) untuk menghasilkan informasi yang berguna. Ekstraksi pengetahuan, analisis pola, penambangan data, dan arkeologi data adalah istilah lain untuk *data Mining*.

Informasi yang dihasilkan dari proses data mining sangat penting bagi sebuah perusahaan atau organisasi sebagai dasar untuk pengambilan keputusan atau tindakan yang tepat di masa depan, dan hasil proses *data mining* bias digunakan untuk memprediksi tren bisnis, dan mengenali pola yang sulit (Mustika et al., 2021).

2.2 Twitter

Twitter merupakan platform jejaring sosial media yang banyak digunakan masyarakat untuk berinteraksi. Twitter diluncurkan pada tahun 2006 di San Fransisco. Jumlah pengguna twitter setiap tahun selalui meningkat dan diperdiksi pengguna twitter indonesia pada tahun 2025 mencapai 18.25 juta pengguna (*Twitter Users in Indonesia 2028 | Statista*, n.d.).

2.3 Python

Python merupakan salah satu bahasa pemrograman yang dinamis, interpretatif, dan terstruktur. Python adalah bahasa pemrograman yang populer dan banyak digunakan karena mudah dipelajari dan dipahami, serta tersedia bermacam-macam library yang dapat mempermudah dalam menyelesaikan berbagai macam keperluan pemrograman.

2.3 JupyterLab

JupyterLab adalah lingkungan pengembangan interaktif (interactive development environment, IDE) yang dikembangkan oleh Jupyter Project. JupyterLab menyediakan interface yang intuitif dan mudah digunakan untuk menulis, mengeksekusi, dan menyajikan kode dan dokumentasi secara bersamaan.

2.4 Analisis Sentimen

Analisis sentimen atau penggalian opini merupakan salah satu bidang ilmu yang menganalisis opini, penilaian, sentimen, emosi dan sikap atas entitas dan atributnya yang diungkapkan dalam teks tertulis (Liu, 2020).

2.5 Naïve Bayes

Naïve Bayes classifiers merupakan metode pengklasifikasian berdasarkan aturan bayes

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

B = *Data dengan class yang belum diketahui*
A = *Hipotesis data B merupakan suatu class spesifik*
P(A|B) = *Probabolitas hipotesis A berdasarkan kondisi B*
P(A) = *Probabilitas hipotesis A*
P(B|A) = *Probabilitas hipotesis B berdasarkan kondisi A*
P(B) = *Probabilitas hipotesis B*

berdasarkan rumus berikut:

Gambar 1 Rumus Naive Bayes

Algoritma *Naïve Bayes* banyak digunakan untuk pengelompokan data dengan jumlah besar dan memprediksi kondisi yang akan terjadi.

3. Metode Penelitian

3.1 Tahapan Penelitian

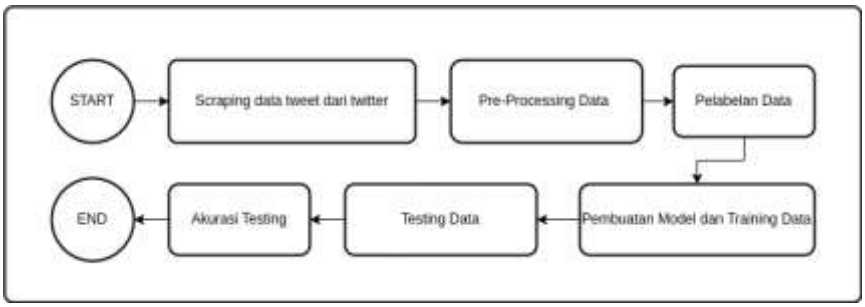
Metodologi yang dipakai untuk mengumpulkan data dan informasi yang diperlukan dalam penelitian ini adalah seperti berikut:

3.1.1 Studi Pustaka

Tahap ini peneliti memasukkan teori-teori yang berhubungan untuk kebutuhan penelitian ini, seperti, pre-processing, pembobotan kata, data mining, data scraping, algoritma naive bayes, analisis sentimen. Peneliti mengumpulkan sumber literatur yang membantu proses penelitian. Literatur tersebut berasal dari jurnal ilmiah, publikasi ilmiah, dan buku.

3.1.2 Pengumpulan Data

Pengumpulan data merupakan tahap kunci dalam penelitian. Teknik pengumpulan data yang benar akan memberikan data yang berkualitas dan sebaliknya. Oleh karena itu, tahap ini perlu dilakukan dengan cermat dan sesuai dengan prosedur dan karakteristik penelitian kualitatif. Konsekuensi dari kesalahan atau kekurangan dalam pengumpulan data bisa sangat merugikan, seperti data yang tidak dapat dipercaya, sehingga hasil penelitian tidak bisa diandalkan. Hasil penelitian seperti itu sangat berbahaya, terutama jika digunakan sebagai dasar pengambilan keputusan politik (Rahardjo, 2011).



Gambar 2 Alur Kerja Penelitian

3.2 Alur Kerja Penelitian

1. Scraping data

   Dalam langkah ini, peneliti melakukan pengambilan tweet melalui teknik *web scraping* yang telah diterangkan pada bagian sebelumnya dari dokumen ini.

2. Pre-Processing

   Dalam proses ini dilakukan pembersihan data tweet menjadi data yang siap diolah oleh mesin. Tahap ini meliputi Case Folding, Stopword, Tokenisasi, dan Stemming. Pada langkah ini peneliti memanfaatkan library NLTK dan Sastrawi untuk melakukan pre-processing.

3. Pelabelan Data

   Setelah data tweet menjadi bersih dilakukan proses pelabelan data. Pelabelan data akan dilakukan secara otomatis dengan membandingkan total kata positif dengan kata negatif yang terdapat pada tweet.

   Jika data tweet mempunyai lebih banyak kata positif maka tweet tersebut akan

dilabeli dengan positif, bila data tweet memiliki lebih banyak kata negatif maka tweet akan dilabeli dengan negatif. Namun jika kata positif dan negatif sama maka dilabeli dengan netral.

4. Pembuatan Model dan Training Data

Di langkah ini tweet yang sudah diberi label akan dibuat beberapa model dengan algoritma Naïve Bayes Gaussian dengan spesifikasi sebagai berikut:

a. Model dengan data latih 90% dan data uji 10%
b. Model dengan data latih 80% dan data uji 20%
c. Model dengan data latih 70% dan data uji 30%
d. Model dengan data latih 60% dan data uji 40%

5. Testing data

Setelah pembuatan model maka dilanjutkan dengan melakukan testing dari masing-masing model yang telah dibuat.

6. Akurasi Testing

Pada tahap ini data hasil testing dilakukan pengecekan akurasi dengan dua metode yaitu percentage split dan cross validation.

4. Hasil dan Pembahasan

4.1 Pengumpulan Data

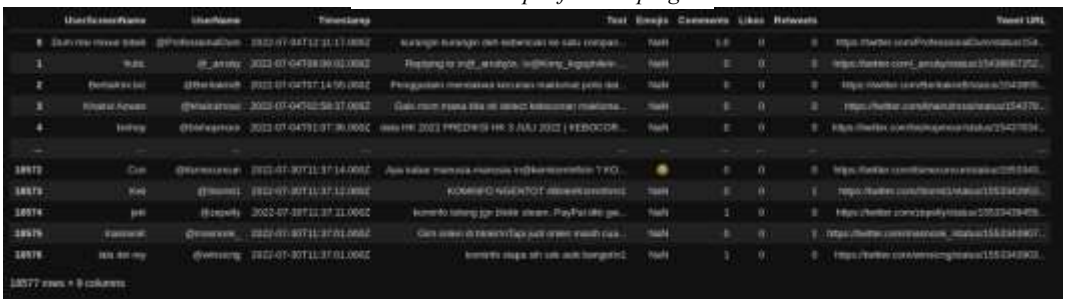
Proses ini dilakukan untuk mengumpulkan data *tweet* dari *Twitter* pada rentang waktu bulan juli 2022 sampai September 2022 yang berkaitan dengan topik kebocoran data seperti ”kebocoran data”, “bjorka” dan “kominfo”. Pengambilan data menggunakan *library* Scweet (*Altimis/Scweet: A Simple and Unlimited Twitter Scraper : Scrape Tweets, Likes, Retweets, Following, Followers, User Info, Images...*, n.d.) dengan perintah sebagai berikut:

```
python scweet.py --words "kebocoran data//bjorka//kominfo" --until 2022-09-30
--since 2022-07-01 --display_type Latest --lang="id" --headless True
```

Gambar 3 Perintah scraping dengan Scweet

Setelah proses scraping selesai maka data akan di simpan di dalam folder output dan akan ada file sesuai dengan data yang kita cari.

Gambar 4 Output file scraping



Dan disini kita mendapat kan 18.577 data tweet dengan contoh data sebagai berikut:

Gambar 5 Contoh data tweet

4.2 Pre-Processing

Pada tahap ini peneliti akan memproses data mentah menjadi data bersih yang agar dapat diproses pada tahap selanjutnya. Berikut adalah tahap yang dilakukan peneliti:

4.2.1 Case Folding

Pada proses ini semua data tweet akan di normalisasi agar menghasilkan data dengan format yang sama (lowercase).

4.2.2 Tokenisasi

Data *tweet* yang telah di *case folding* akan dipotong-potong per kata yang disebut *token*. Tahap ini juga bertujuan menghilangkan tanda baca seperti @/!#$% dan merubah kata slang atau kata tidak baku ke ke bentuk bakunya contoh nya “met” menjadi “selamat”.

Tahap ini juga melakukan proses stopword yaitu penghapusan kata-kata yang tidak memiliki arti penting seperti “musuh”, “masyarakat”, “ongkos" dan sebagainya.

4.2.3 Stemming

Stemming merupakan proses untuk menghilangkan imbuhan pada suatu kata (Guterres et al., 2019). Misalnya kata “mendengarkan”, “dengarkan”, “didengarkan” akan ditransformasi menjadi kata “dengar”.

4.3 Frekuensi Data

Kemudian data tweet yang telah melalui pre-processing bisa kita coba untuk mencari kata yang sering muncul.

1. Pertama memecah semua data tweet menjadi perkata kedalam satu variable all_words.

2. Kemudian menggunakan bantuan library nltk untuk mencari frekuensi kata dari dalam variable all_words dengan fungsi nltk.FreqDist().
3. Memunculkan 10 kata teratas frekuensi dengan fungsi word_dist.most_common().

```
[('kominfo', 16574),
 ('data', 6842),
 ('kemkominfo', 5602),
 ('blokir', 5390),
 ('bocor', 5149),
 ('daftar', 3455),
 ('pse', 3089),
 ('jul', 3042),
 ('indonesia', 2823),
 ('digital', 2717)]
```

Gambar 6 Frekuensi Data

Disini kata “kominfo” menjadi yang paling tinggi dengan frekuensi muncul sebanyak 16.574 kali.

Gambar 7 Word Cloud Frekuensi Data

4.4 Pelabelan

Pada tahap ini tiap *tweet* yang telah bersih akan diberi label secara otomatis apakah *tweet* tersebut positif, negatif atau netral. Dengan contoh *tweet* “Kominfo harusnya fokus ngurusin kebocoran data pribadi dan jadi lembaga yang berpihak pada korbannya. Konten-konten non-konsensual masih gampang banget diakses lewat internet, dan berdampak banget sama pembunuhan karakter seseorang. Dalam kasus kayak gini, Kominfo ke mana?” terdapat 0 data positif dengan 1 data negatif sehingga perhitungan skor *tweet* dapat menggunakan rumus berikut:

Skor = (Σ kata positif)- (Σ kata negatif)

Contoh perhitungan:

Skor = (Jumlah kata positif) – (Jumlah kata negatif)

Skor = 0 – 1 = -1

Apabila hasil akhir dari perhitungan memperoleh skor -1 atau kurang dari 0, maka dapat dikategorikan sebagai tweet bernuansa negatif..

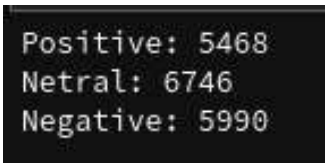
Dari perhitungan diatas bisa digunakan untuk melakukan pelabelan secara otomatis dengan kriteria

Data Positif = Skor > 0

Data Netral = Skor == 0

Data Negatif = Skor < 0

Dari kriteria diatas maka mendapat 5.468 tweet Positif, 6.746 tweet Netral dan 5.990



Gambar 8 Total Sentimen

tweet Negatif.

4.5 Pengujian

Pengujian dilakukan menggunakan metode *percentage split* dan *cross validation* untuk mengetahui akurasi pada setiap model data uji. Pengujian direalisasikan dengan kode python berikut:

```
[8] def checkAccuracy(x, y, x_test, y_test, filename, reTrain = False):
      isModelExist = path.isfile("{}/Model/{}.pickle".format(LOCATION, filename))
      gnb = GaussianNB()
      if not reTrain and isModelExist:
        f = open("{}/Model/{}.pickle".format(LOCATION, filename), 'rb')
        gnb = pickle.load(f)
        f.close()
      else:
        gnb.fit(x, y)
      prediction = gnb.predict(x_test)
      accuracy = accuracy_score(y_test, prediction)
      scores = cross_val_score(gnb, x_test, y_test, cv = 10, scoring='accuracy')
      print(classification_report(y_test, prediction))
      ConfusionMatrixDisplay.from_predictions(y_test, prediction)
      plt.show()
      f = open("{}/Model/{}.pickle".format(LOCATION, filename), 'wb')
      pickle.dump(gnb, f)
      f.close()
      return accuracy * 100, scores, prediction
```

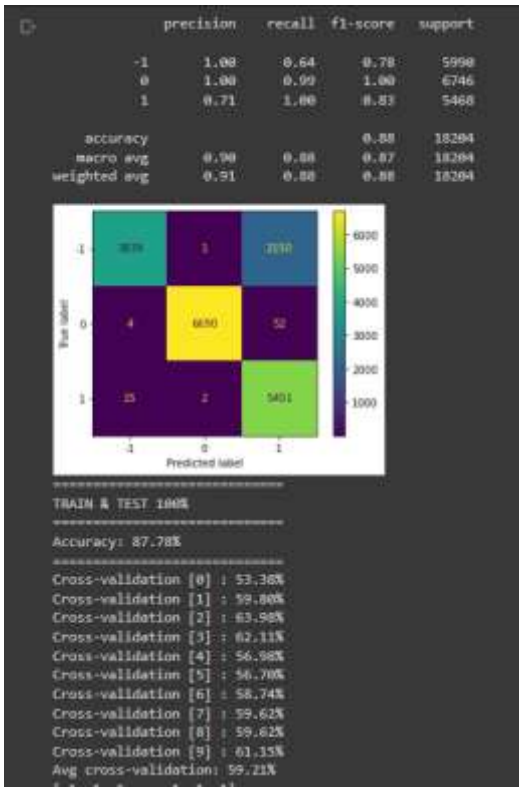
Gambar 9 Fungsi training dan pengecekan akurasi

Dari kode diatas akan menghasilkan keluaran akurasi seperti berikut:

4.5.1 Data Train 90% dan Data Test 10%

Untuk pengujian pertama, peneliti membagi 90% data sebagai data latih dan 10% sebagai data uji. Setelah melakukan pengujian, hasil akurasi yang diperoleh sebesar 66.28%. Hasil akurasi pengujian cross-validation juga dapat ditemukan sebagai berikut:

CV	Akurasi
1	61.20%
2	57.69%
3	60.99%
4	59.89%
5	51.10%
6	63.19%



Gambar 10 Output Fungsi Training dan Pengecekan Akurasi

7	58.24%
8	59.62%
9	59.62%
10	61.15%
Avg	59.21%

Table 1 Akurasi Train 90% dan Test 10%

Berdasarkan hasil cross validation diatas mendapat nilai rata-rata 58.76% dengan tingkat akurasi paling tinggi pada validasi ke-6 dengan nilai 63.19% dan paling rendah

pada validasi ke-5 dengan nilai akurasi 51.10%.

4.5.2 Data Train 80% dan Data Test 20%

Untuk pengujian kedua, peneliti membagi 80% data sebagai data latih dan 20% sebagai data uji. Setelah melakukan pengujian, hasil akurasi yang diperoleh sebesar 67.01%. Hasil akurasi pengujian cross-validation juga dapat ditemukan sebagai berikut:

CV	Akurasi
1	66.03%
2	60.99%
3	61.26%
4	62.91%
5	63.46%
6	66.21%
7	66.76%
8	67.58%
9	62.64%
10	63.74%
Avg	64.16%

Table 2 Akurasi Train 80% dan Test 20%

Berdasarkan hasil cross validation diatas mendapat nilai rata-rata 64.16% dengan tingkat akurasi paling tinggi pada validasi ke-8 dengan nilai 67.58% dan paling rendah pada validasi ke-2 dengan nilai akurasi 60.99%.

4.5.3 Data Train 70% dan Data Test 30%

Hasil pengujian ketiga menunjukkan bahwa peneliti menggunakan 70% data untuk latih dan 30% untuk uji. Hasil akurasi yang diperoleh adalah 66.18%, dengan akurasi cross-validation seperti ini:

CV	Akurasi
1	66.54%
2	64.90%
3	63.00%
4	65.57%
5	68.32%
6	68.68%
7	65.20%
8	65.93%
9	66.48%
10	67.58%
Avg	66.22%

Table 3 Akurasi Train 70% dan Test 30%

Berdasarkan hasil cross validation diatas mendapat nilai rata-rata 66.22% dengan tingkat akurasi paling tinggi pada validasi ke-6 dengan nilai 68.68% dan paling rendah pada validasi ke-3 dengan nilai akurasi 63.00%.

4.5.4 Data Train 60% dan Data Test 40%

Pada pengujian keempat, sebanyak 60% data digunakan sebagai data latih dan 40% sebagai data uji. Hasil pengujian menunjukkan akurasi 66.29%, dengan akurasi cross-validation sebagai berikut:

CV	Akurasi
1	65.71%
2	62.96%
3	64.56%
4	65.93%
5	68.41%
6	66.48%
7	64.15%
8	62.64%
9	64.42%
10	65.80%
Avg	65.11%

Table 4 Akurasi Train 60% dan Test 40%

Berdasarkan hasil cross validation diatas mendapat nilai rata-rata 65.11% dengan tingkat akurasi paling tinggi pada validasi ke-5 dengan nilai 68.41% dan paling rendah pada validasi ke-8 dengan nilai akurasi 62.64%.

5. Kesimpulan

Berdasarkan hasil pengujian yang dilakukan, peneliti dapat mengambil kesimpulan dari penelitian sebagai berikut:

a. Sentimen Masyarakat
   Sentimen masyarakat memiliki dominasi sentimen positif. Persentase berkisar antara 46,7% - 50%. Sehingga dapat disimpulkan bahwa masyarakat masih mempercayai pemerintah terkait isu kebocoran data.
b. Hasil Akurasi
   1. Hasil akurasi pada pengujian stabil antara 66.18% sampai 67.01%
   2. Peningkatan jumlah data uji tidak selalu meningkatkan tingkat akurasi model.

Referensi

Altimis/Scweet: A simple and unlimited twitter scraper: Scrape tweets, likes, retweets, following, followers, user info, images... (n.d.). Retrieved January 17, 2023, from https://github.com/Altimis/Scweet

Guterres, A., Gunawan, & Santoso, J. (2019). Stemming Bahasa Tetun Menggunakan Pendekatan Rule Based. Teknika, 8(2), Article 2. https://doi.org/10.34148/teknika.v8i2.224

Liu, B. (2020). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, ii–ii.

Mustika, M. (Mustika), Ardilla, Y. (Yunita), Manuhutu, A. (Abraham), Ahmad, N. (Nazaruddin), Hasbi, I. (Imanuddin), Guntoro, G. (Guntoro), Manuhutu, M. A. (Melda), Ridwan, M. (Mohamad), Hozairi, H. (Hozairi), Wardhani, A. K.

(Anindya), Alim, S. (Syariful), Romli, I. (Ikhsan), Religia, Y. (Yoga), Octafian, D. T. (D), Sufandi, U. U. (Unggul), & Ernawati, I. (Iin). (2021). DATA MINING DAN APLIKASINYA. https://repository.penerbitwidina.com/publications/351768/

Rahardjo, M. (2011). Metode pengumpulan data penelitian kualitatif.

Rozi, I. F., Pramono, S. H., & Dahlan, E. A. (2012). Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. Jurnal EECCIS (Electrics, Electronics, Communications, Controls, Informatics, Systems), 6(1), 37–43. https://doi.org/10.21776/JEECCIS.V6I1.164

S, D. (n.d.). Countries with most Twitter users 2022. Statista. Retrieved September 14, 2022, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Twitter users in Indonesia 2028 | Statista. (n.d.). Retrieved January 3, 2023, from https://www.statista.com/forecasts/1145550/twitter-users-in-indonesia

Zaskya, M., Boham, A., Jackelin, L., & Lotulung, H. (2021). Twitter Sebagai Media Mengungkapkan Diri Pada Kalangan Milenial. ACTA DIURNA KOMUNIKASI, 3(1). https://ejournal.unsrat.ac.id/v3/index.php/actadiurnakomunikasi/article/view/31979

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/)