Twitter Sentiment Analysis Menggunakan Bahasa Indonesia

Nama Anggota

2301937173 - Dean Ananda Ramadhan
2301853123 - Leonardo Ignatius
2301873182 - Matthew Liem
2301855154 - Axel Lie

Latar Belakang

Di era digital ini informasi merupakan hal yang lazim digunakan mulai dari individu sampai perusahaan internasional. Hal ini, tentu saja membawa kemajuan di beberapa sektor. Namun, dengan kemajuan tersebut muncul juga sebuah kekurangan.

Salah satu kekurangannya adalah ujaran kebencian atau yang sering disebut hate speech. Ujaran kebencian dapat memiliki berbagai bentuk seperti sarkasme atau sindiran. Hal ini dapat membahayakan beberapa hal atau bahkan sekelompok orang di luar konteks. Contoh lain yang dapat merusak komunitas adalah hoax atau kabar bohong.

Dengan kombinasi kelemahan seperti itu dan kecepatan penyebaran informasi ini dapat menyebabkan kekerasan, reputasi buruk, dan cyberbullying. Salah satu contoh dari hal tersebut adalah #chinesevirus, beberapa bulan setelah tagar ini menjadi trending di twitter sebagian besar orang asia. Keturunan orang asia mendapat serangan verbal beberapa bahkan serangan fisik.

Dengan perkembangan informasi yang pesat dibutuhkan alat untuk memilah dan memantau tweet Twitter dengan cepat dan akurat. Kami berupaya untuk mengatasi masalah ini dengan membuat proyek sentiment analysis twitter dengan bantuan AI. Dengan proyek ini, kita dapat memantau dan mempolarisasikan tweet yang mengarah ke negatif atau positif.

Disisi lain, AI yang digunakan sentiment analysis menggunakan Bahasa Indonesia masih sangat sedikit dan terbatas. Oleh karena itu, topik ini diambil juga sebagai percobaan untuk meningkatkan NLP menggunakan Bahasa Indonesia.

Tujuan dan Manfaat

Tujuan :

- Menentukan sentimen baik/buruk suatu tren yang terjadi saat ini dengan menggunakan analisis berdasarkan data dari media sosial Twitter.
- Menemukan insight atas data yang akan dianalisa dengan NLP.
- Mencapai minimal 80% akurasi, recall, presisi, dan f1 score untuk model.

Manfaat:

1. Social media monitoring

Media sosial menjadi hal yang tidak terpisahkan dari sebuah bisnis. Baik marketing, sales, hingga branding dilakukan melalui jenis media yang satu ini. Tak hanya itu, publik dan pelanggan juga lekat dengan opini di media sosial. Oleh karena itu, penting untuk mengetahui sentimen apa yang dibicarakan publik terhadap bisnismu. Kamu bisa menilai apakah mereka banyak membicarakan hal positif, negatif, atau netral tentang acara-tv nya.

2. Customer feedback

Jika social media dan brand monitoring dilakukan untuk melihat sentimen publik secara luas, tidak dengan customer feedback. Customer feedback bertujuan untuk mengumpulkan pendapat dari pelanggan. Kamu bisa melakukannya dengan metode survei dan menganalisisnya dengan teknik sentiment analysis.

3. Market research

Hal yang tidak kalah penting dalam suatu bisnis adalah market research atau riset pasar. User bisa memanfaatkan sentiment analysis untuk menilai apa yang sedang disukai dan tidak disukai market.

Data

Data yang digunakan berasal dari github user “rizalespe” yang berjudul dataset_tweet_sentimen_tayangan_tv.csv dengan jumlah entry mencapai 400 entries.

Link :
https://github.com/rizalespe/Dataset-Sentimen-Analisis-Bahasa-Indonesia/blob/master/dataset_tweet_sentimen_tayangan_tv.csv

Fitur dari dataset di atas berisikan: Id, Sentiment, Acara TV, Jumlah Retweet, dan Text Tweet.

Fitur	Data Type

Id	int
Sentiment	String
Acara TV	String
Jumlah Retweet	int
Text Tweet	String

Metodologi Penelitian

Tahapan dalam melakukan analisis sentimen dengan beberapa metode klasifikasi SVM, MNB, dan Neural network dimulai dengan input data yang berupa data latih dan data uji yang kemudian diproses pada tahapan pre-processing hingga proses klasifikasi untuk menentukan prediction class.

Pre-processing merupakan tahapan awal yang akan dilalui dalam memproses teks. Langkah-langkah pre-processing :

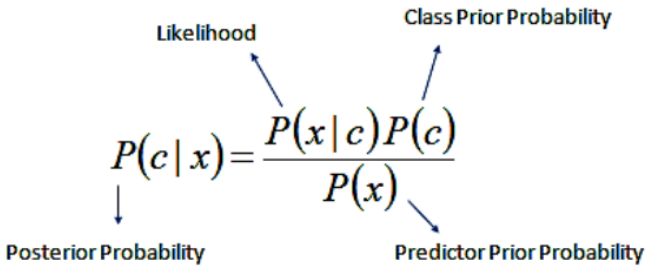
1. Label Encoding
2. Case Folding
3. Text Cleaning :
    1) Membersihkan @ (mention)
    2) Membersihkan # (hashtag)
    3) Membersihkan https:// dan http://
4. Translation to English
5. Stop Word Removal
6. TF-IDF
7. Feature Engineering

Setelah melakukan pre-processing, kami langsung membuat model untuk klasifikasi prediction class :

1. Multinomial Naive Bayes

Metode ini memanfaatkan teorema probabilitas yaitu teorema bayes dan fungsionalitas data mining yaitu klasifikasi naive bayesian. Metode Multinomial Naive Bayes merupakan algoritma yang naïve karena mengasumsikan independensi diantara kemunculan kata-kata dalam dokumen, tanpa memperhitungkan urutan kata dan informasi konteks dalam kalimat atau dokumen secara umum. Multinomial Naive bayes adalah salah satu metode bayes yang dipakai dengan memperhitungkan frekuensi masing-masing

kemunculan kata dalam sebuah dokumen dan probabilitas. Kelebihan naive bayes multinomial diantaranya adalah tingkat akurasi yang tinggi, mudah diimplementasikan, waktu komputasi yang rendah serta error rate yang minimum. Multinomial Naïve bayes dapat menangani ukuran kosakata dalam jumlah besar serta mereduksi tingkat error



Persamaan:

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

2. Support Vector Machine

SVM adalah metode supervised learning yang menganalisis data dan mengenali pola yang digunakan untuk klasifikasi. Dalam pemodelan klasifikasi, SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya. SVM juga dapat mengatasi masalah klasifikasi dan regresi dengan linier maupun non linear.

SVM digunakan untuk mencari hyperplane terbaik dengan memaksimalkan jarak antar kelas. Hyperplane adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas. Dalam 2-D fungsi yang digunakan untuk klasifikasi antar kelas disebut sebagai line whereas, fungsi yang digunakan untuk klasifikasi antas kelas dalam 3-D disebut plane similarly, sedangkan fungsi yang digunakan untuk klasifikasi di dalam ruang kelas dimensi yang lebih tinggi disebut hyperplane.

Persamaan Support Vector Machine:

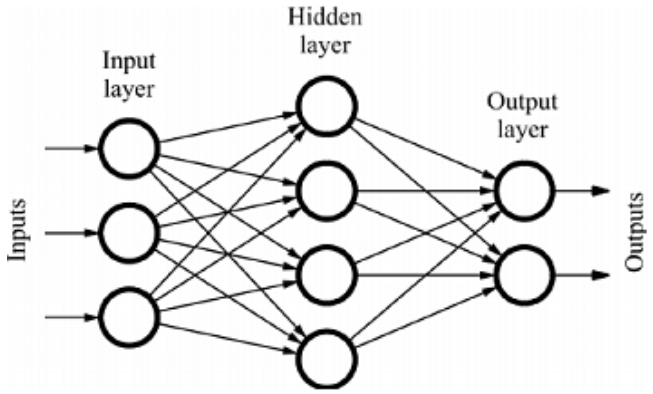
$f(x) = w.\ x + b$ (4) atau $f(x) = \sum aiyiK(x, xi\ ) + b\ m\ i$=1

Keterangan :

w : parameter hyperplane yang dicari (garis yang tegak lurus antara garis hyperplane dan titik support vector)
x : titik data masukan Support Vector Machine ai : nilai bobot setiap titik data $K(x, xi\ )$: fungsi kernel
b : parameter hyperplane yang dicari (nilai bias)

3. Feed-forward Neural Network

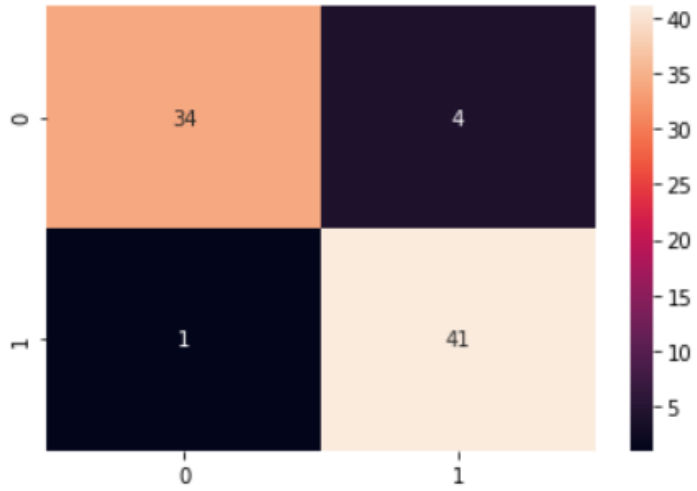
Feed-forward Neural Network adalah neural network yang mempunyai tiga layer yaitu layer input, layer tersembunyi (hidden layer) dan layer output. Ada beberapa struktur model dari jaringan syaraf tiruan dan salah satunya adalah jaringan syaraf tiruan berstruktur feedforward (feedforward neural network). Secara sederhana, jaringan syaraf tiruan yang mempunyai struktur feedforward mempunyai karakteristik tidak ada pengulangan pembelajaran (loop) di mana signal bergerak dari layer input dan melewati layer tersembunyi dan kemudian menuju layer output. Keuntungan dari neural network terletak pada aspek teoritis berikut. Pertama, jaringan saraf adalah metode self-adaptive yang didorong oleh data karena mereka dapat menyesuaikan diri dengan data tanpa eksplisit spesifikasi bentuk fungsional atau distribusi untuk model yang mendasarinya. Kedua, mereka berfungsi universal yang mendekati bahwa neural network dapat memperkirakan fungsi apa pun dengan akurasi sewenang-wenang.


Hidden
layer
Input
layer
Output
layer
Inputs
Outputs

Hasil Eksperimen dan Analisis

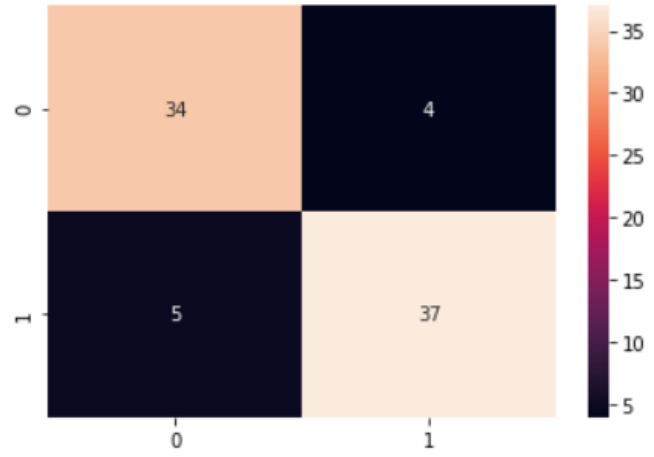
Multinomial Naive Bayes:

```
Acurracy: 93.75%
Recall: 97.62%
Precision: 91.11%
F1 Score: 94.25%
```



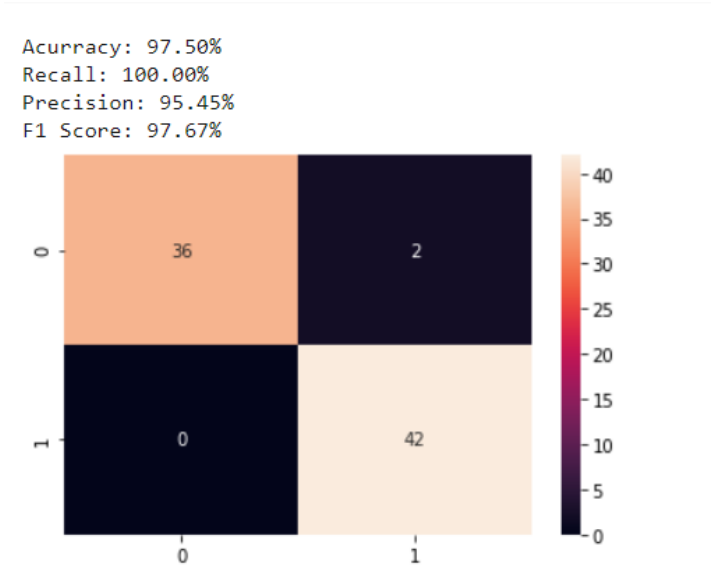
Support Vector Machine:

```
Acurracy: 88.75%
Recall: 88.10%
Precision: 90.24%
F1 Score: 89.16%
```

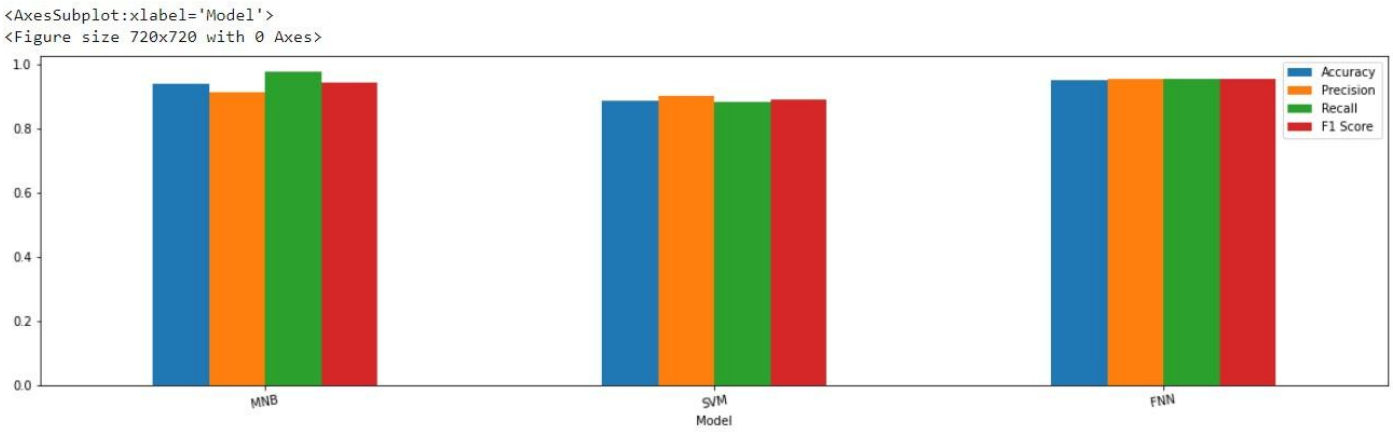


FNN:

Training MSE : 3.3436e-04 - Training Accuracy: 1.0



Bar Graph Untuk Membandingkan Accuracy , Precision, Recall, F1 Score pada Testing data:



Simpulan

Dalam laporan ini kami menampilkan hasil eksperimen tentang analisa sentimen pada sosial media twitter spesifik pada topik acara pada stasiun televisi. motivasi utama dari percobaan ini untuk membuat algoritma yang dapat mengetahui sentimen dari suatu acara televisi yang beredar di pasaran menurut pengguna sosial media Twitter dengan menggunakan Supervised Machine Learning. Kita mendapatkan hasil evaluasi prediksi yang baik dengan menggunakan model machine learning klasifikasi seperti *Support Vector Machine, Multinomial Naive Bayes.* dan *Feed Forward Artificial Neural Network*. Setelah melakukan percobaan dengan menggunakan ketiga model tersebut didapatkan akurasi terbaik dari model *Feedforward*

Artificial Neural Network dengan akurasi tertinggi 97.50%. Dengan begitu sampai di akhir percobaan kami. Masih terdapat beberapa percobaan tambahan yang dapat meningkatkan hasil prediksi percobaan ini dan dapat dilanjutkan di masa yang akan datang.

Referensi

1. Rofiqoh, Umi & Perdana, Rizal & Fauzi, Muhammad. (2017). Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter dengan Metode Support Vector Machine dan Lexicon Based Features. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK). 1. 1725-1732.

2. Firmansyah, Ro'i & Fauzi, Muhammad & Afirianto, Tri. (2016). SENTIMENT ANALYSIS PADA REVIEW APLIKASI MOBILE MENGGUNAKAN METODE NAÏVE BAYES DAN QUERY EXPANSION. DORO PTIIK. 8.

3. P. Borele and D. Borikar, “An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques,” vol. 18, no. 2, pp. 64–69, 2016, doi: 10.9790/0661-1802056469.