A blurred background image showing two people from behind, both wearing light-colored shirts and dark trousers, sitting at a desk and looking at a laptop screen together.

AI Career Bootcamp

# Text Classification using RNN & LSTM

## Pembukaan

Guide Book ini memuat beberapa informasi-informasi utama yang akan disampaikan oleh mentor di program AI Career Bootcamp yang bisa dijadikan pegangan para students untuk mempersiapkan diri sebelum sesi Live Class berlangsung.

## Deskripsi

Text classification, juga dikenal sebagai klasifikasi teks, adalah tugas dalam pemrosesan bahasa alami (natural language processing) yang bertujuan untuk mengelompokkan teks ke dalam kategori atau kelas yang telah ditentukan sebelumnya. Tujuan utama dari text classification adalah untuk mengenali pola, tema, atau topik tertentu dalam teks dan mengatributkannya ke kelas yang sesuai.

Proses text classification melibatkan penggunaan teknik pembelajaran mesin atau statistik untuk mempelajari pola dari teks berlabel yang sudah ada. Pendekatan yang umum digunakan termasuk menggunakan algoritma Machine Learning seperti Naive Bayes, Logistic Regression, Support Vector Machines (SVM) termasuk algoritma Deep Learning seperti Recurrent Neural Networks (RNN) dan Long Short-term Memory (LSTM).

Langkah-langkah umum dalam text classification adalah sebagai berikut:

- 1 Pra-pemrosesan teks: Teks biasanya melalui tahap pra-pemrosesan untuk menghilangkan karakter khusus, mengubah teks menjadi huruf kecil, menghapus stop words (kata-kata yang umum dan tidak memberikan banyak informasi), melakukan stemming atau lemmatisasi, dan menerapkan teknik normalisasi lainnya untuk membersihkan dan merapikan teks.
- 2 Fitur extraction: Fitur-fitur yang relevan perlu diekstraksi dari teks untuk memberikan representasi numerik yang dapat digunakan oleh algoritma pembelajaran mesin. Beberapa metode ekstraksi fitur yang umum meliputi penghitungan jumlah kata, penggunaan TF-IDF, n-gram, atau pemodelan bahasa seperti Word2Vec atau GloVe.
- 3 Pembuatan model: Setelah fitur-fitur diekstraksi, model pembelajaran mesin dibangun dengan menggunakan data latih (labeled data). Model dapat berupa model klasifikasi seperti Naive Bayes, SVM, atau model deep learning seperti RNN, LSTM yang dilatih dengan algoritma seperti backpropagation.

- 4 Evaluasi model: Model yang dibangun dievaluasi dengan menggunakan data uji (test data) untuk mengukur kinerjanya. Metrik evaluasi yang umum digunakan dalam text classification meliputi akurasi (accuracy), presisi (precision), recall, dan F1-score.
- 5 Prediksi: Setelah model dievaluasi dengan baik, model tersebut dapat digunakan untuk melakukan prediksi pada teks yang belum dilihat sebelumnya. Teks baru akan diklasifikasikan ke dalam kelas yang sesuai berdasarkan pembelajaran yang telah dilakukan oleh model.

Text classification memiliki berbagai aplikasi dalam pemrosesan bahasa alami, seperti analisis sentimen, kategorisasi berita, deteksi spam, analisis topik, dan banyak lagi. Dengan kemampuan untuk mengklasifikasikan teks secara otomatis, text classification memungkinkan pemrosesan dan analisis efisien dalam skala besar dan mendukung berbagai aplikasi yang bergantung pada pemahaman dan pengelompokan teks.

## Mengenal RNN

Recurrent Neural Network (RNN) adalah jenis algoritma dalam pembelajaran mesin yang terkenal karena kemampuannya dalam memproses sequential data atau data berurut seperti data teks. RNN memiliki struktur berulang yang memungkinkannya menggunakan informasi dari langkah waktu sebelumnya saat memproses langkah waktu saat ini. Dalam RNN, setiap simpul memiliki hidden state yang berfungsi sebagai memori internal yang dapat menyimpan informasi kontekstual dari urutan sebelumnya. Hal ini memungkinkan RNN untuk mengenali pola dan ketergantungan temporal dalam data berurutan.

Salah satu kelemahan RNN tradisional adalah kesulitan dalam mempertahankan informasi jangka panjang dalam urutan yang panjang. Untuk mengatasi masalah ini, dikembangkan variasi RNN yang lebih canggih seperti Long Short-Term Memory (LSTM). LSTM menggunakan pintu (gate) untuk mengatur aliran informasi, mengendalikan kapan dan bagaimana informasi dapat diingat/dilupakan.

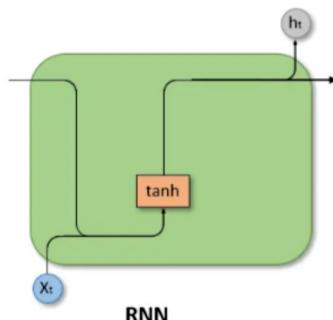


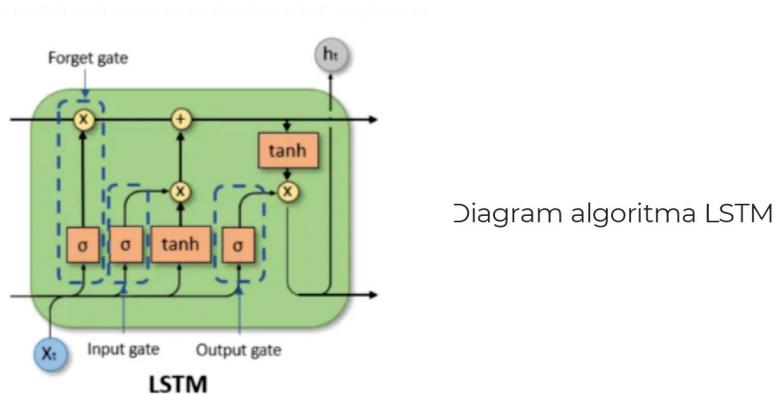
Diagram algoritma RNN

RNN dan variasi LSTM telah diterapkan dalam berbagai aplikasi. Misalnya, dalam pemrosesan bahasa alami, RNN dapat digunakan untuk pemodelan bahasa, terjemahan mesin, analisis sentimen, dan generasi teks. Dalam pemrosesan deret waktu, RNN berguna untuk prediksi deret waktu, analisis pola, dan pengenalan gerakan. Algoritma RNN memungkinkan mesin untuk mempelajari dan memahami informasi temporal dalam data berurutan, membuatnya menjadi alat yang kuat dalam memproses data yang terstruktur secara temporal dan mendapatkan wawasan yang lebih baik dari urutan data.

## Mengenal LSTM

Long Short-Term Memory (LSTM) adalah sebuah variasi dari Recurrent Neural Network (RNN) yang dirancang khusus untuk mengatasi masalah mempertahankan informasi jangka panjang dalam sequential data atau data berurut seperti data teks. LSTM menjadi populer dalam pemrosesan bahasa alami dan tugas pemodelan deret waktu. Salah satu keunggulan utama LSTM adalah kemampuannya untuk mengelola informasi kontekstual yang berjalan lama dan mempertahankannya melalui pintu (gate) khusus.

LSTM mencapai hal ini dengan menggunakan tiga pintu utama: pintu lupa (forget gate), pintu input (input gate), dan pintu keluaran (output gate). Pintu lupa memungkinkan LSTM untuk memutuskan informasi apa yang harus dihapus dari hutan belakang (hidden state) berdasarkan informasi dari langkah waktu sebelumnya. Pintu input memungkinkan LSTM untuk memutuskan informasi apa yang harus diingat dari input saat ini dan menggabungkannya dengan hutan belakang. Pintu keluaran memungkinkan LSTM untuk memutuskan informasi apa yang harus dihasilkan dari hutan belakang yang telah diperbarui.



Dengan menggunakan pintu-pintu ini, LSTM dapat mempelajari ketergantungan jarak panjang dalam urutan data. Hal ini menjadikan LSTM sangat efektif dalam memproses teks, seperti pemodelan bahasa, analisis sentimen, dan terjemahan mesin. Selain itu, dalam pemodelan deret waktu, LSTM dapat digunakan untuk prediksi, deteksi anomali, dan pemrosesan sinyal. Keunggulan LSTM dalam mempertahankan informasi jangka panjang dan mengelola ketergantungan temporal menjadikannya pilihan populer dalam berbagai aplikasi yang melibatkan data berurutan.