Laporan Proyek Natural Language Processing (NLP)

**KLASIFIKASI KATEGORI BERITA MENGGUNAKAN MODEL LOGISTIC REGRESSION**

A blue logo with green and blue text

Description automatically generated

Dosen Pengampu:

Inna Sabily Karima S.Kom., M.Kom.

Disusun oleh:

David Hasiholan Nadeak (41521010099)

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS MERCU BUANA

TAHUN 2024

**KATA PENGANTAR**

Segala puji syukur hanya milik Tuhan yang Maha Esa. Atas berkat rahmat dan karuniaNya, penulis dapat menyelesaikan laporan proyek ini dengan judul “KLASIFIKASI KATEGORI BERITA MENGGUNAKAN MODEL LOGISTIC REGRESSION”.

Laporan proyek ini disusun untuk memberikan dokumentasi tentang klasifikasi kategori berita menggunakan model Logistic Regression. Isi laporan proyek ini disajikan dalam latar belakang, metodologi, hasil dan analisis, serta kesimpulan guna memudahkan pembaca dalam memahami isi laporan proyek ini secara menyeluruh.

Dalam penulisan laporan proyek ini, penulis menyadari bahwa masih banyak kekurangan yang perlu perbaikan, kritik yang membangun sangat penulis harapkan dari dosen pengampu mata kuliah Natural Language Processing (NLP). Juga usulan dari para pembaca laporan proyek ini, semoga laporan proyek ini dapat diperkaya melalui evaluasi terus menerus. Akhir kata, penulis berharap semoga laporan proyek ini dapat memberikan manfaat bagi kita semua.

Jakarta, 15 Desember 2024

David Hasiholan Nadeak

**LATAR BELAKANG**

Dalam era digitalisasi dan perkembangan teknologi informasi yang pesat, jumlah data yang dihasilkan, termasuk data teks, meningkat secara signifikan setiap harinya. Salah satu bentuk data teks yang paling umum adalah berita yang dipublikasikan melalui berbagai platform digital seperti situs web, media sosial, dan aplikasi berita. Berita memainkan peran penting dalam menyampaikan informasi, namun lonjakan volume berita dapat menimbulkan tantangan bagi editor dalam mengorganisir dan mengelompokkan berita yang hendak dipublikasikan dengan topik yang relevan. Oleh karena itu, diperlukan sistem klasifikasi otomatis yang mampu mengelompokkan berita ke dalam kategori yang sesuai.

Natural Language Processing/NLP merupakan cabang dari kecerdasan buatan (Artificial Intelligence/AI) yang berfokus pada interaksi antara komputer dan bahasa manusia (natural language). NLP pada penelitian ini digunakan untuk mengklasifikasikan berita ke dalam beberapa kategori berdasarkan isi dari teks tersebut. Proses klasifikasi ini dapat membantu pengguna dalam memperoleh informasi yang lebih terstruktur, cepat, dan akurat.

Adapun dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari [Kaggle: World News Category](https://www.kaggle.com/datasets/khoshbayani/news-texts). Dataset ini memiliki kolom “date”, “text”, dan “label”. Dimana “text” merupakan fitur dan “label” merupakan target. Setiap baris data diberi label dengan kelas “politic”, “technology”, dan “science”.

Dalam penelitian ini, model Logistic Regression digunakan untuk membangun sistem klasifikasi kategori berita. Logistic Regression merupakan salah satu metode machine learning yang populer dalam tugas klasifikasi biner maupun multikelas karena kesederhanaannya, efisiensi komputasi, dan kemampuannya memberikan hasil yang cukup baik pada dataset dengan dimensi tinggi seperti data teks.

Selain itu, penelitian ini juga menerapkan dua teknik representasi teks, yaitu TF-IDF (Term Frequency-Inverse Document Frequency) dan Bag of Words, untuk mengetahui teknik mana yang lebih baik dalam meningkatkan kinerja model klasifikasi. Dengan menerapkan Logistic Regression bersama kedua teknik tersebut, penelitian ini bertujuan untuk membangun model klasifikasi yang efektif dan efisien dalam mengklasifikasikan kategori berita berdasarkan teks yang tersedia.

Melalui penelitian ini, diharapkan dapat dihasilkan suatu sistem yang tidak hanya mendukung pengelompokan berita secara otomatis, tetapi juga memberikan kontribusi bagi pengembangan teknologi klasifikasi teks berita. Selain itu, hasil dari penelitian ini diharapkan dapat membantu industri media dan pengembang sistem informasi dalam menyajikan berita yang lebih terstruktur dan terorganisir kepada pembaca.

Rumusan Masalah:

1. Bagaimana performa model Logistic Regression dalam mengklasifikasikat kategori berita berdasarkan teknik representasi teks TF-IDF dan Bag of Words?
2. Bagaimana distribusi kelas dataset berdasarkan label (original) dan hasil klasifikasi oleh model Logistic Regression?

Tujuan Penelitian:

1. Menganalisis performa model Logistic Regression dalam mengklasifikasikan kategori berita berdasarkan teknik representasi teks TF-IDF dan Bag of Words.
2. Mengidentifikasi distribusi kelas dataset berdasarkan label asli (original) dan hasil klasifikasi oleh model Logistic Regression.

**METODOLOGI**

A screenshot of a computer

Description automatically generated

Gambar 1. Metode Penelitian

1. Pengumpulan Data

Tahap awal penelitian dilakukan dengan mengumpulkan dataset berupa file news.csv yang berisi dua kolom utama, yaitu text sebagai fitur dan label sebagai target berita. Kolom label memiliki kelas seperti politic, technology, dan science yang akan diklasifikasikan dalam penelitian ini. Nantinya dataset akan dipreprocessing agar siap digunakan untuk training model.

1. Preprocessing Data

Langkah pertama adalah menghapus data yang memiliki nilai kosong pada kolom text atau label. Selanjutnya, teks pada kolom text dinormalisasi dengan mengubah semua huruf menjadi kecil (lowercase), menghapus simbol, angka, dan karakter khusus, serta menghapus stopwords bahasa Inggris. Selain itu, kelas yang berada di kolom label dikonversi ke dalam format numerik (0: ‘politic’, 1:’science’, 2:’technology’). Semua tahapan preprocessing ini dilakukan agar data siap digunakan sebagai data training dan data testing.

1. Split Data

Data yang telah diproses kemudian dibagi menjadi dua bagian, yaitu data latih (80%) dan data uji (20%). Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengevaluasi performa model.

1. Vectorization

Teks pada kolom text yang telah diproses dikonversi ke dalam bentuk numerik menggunakan teknik TF-IDF dan Bag of Words. Representasi numerik ini memungkinkan model machine learning untuk memahami data teks.

1. Modeling

Model klasifikasi yang digunakan dalam penelitian ini adalah Logistic Regression. Model dilatih menggunakan data training yang telah di-vectorize untuk mempelajari pola teks dan hubungan antara fitur teks dan label. Setelah proses pelatihan selesai, model diuji menggunakan data testing untuk menghasilkan prediksi kategori berita.

1. Evaluasi Model

Kinerja model dievaluasi menggunakan berbagai metrik performa, seperti akurasi, precision, recall, dan F1-score. Tahapan ini penting untuk dilakukan karena berfungsi untuk memberikan gambaran performa atau kinerja model dalam melakukan klasifikasi data.

**HASIL DAN ANALISIS**

**3.1. Klasifikasi Menggunakan Teknik TF-IDF**

TF-IDF digunakan sebagai teknik representasi teks digunakan sebagai teknik representasi teks untuk mengubah data teks menjadi bentuk numerik sehingga dapat membantu model Logistic Regression dalam proses klasifikasi. TF-IDF berfungsi untuk memberikan bobot pada setiap kata berdasarkan frekuensi kemunculannya dalam sebuah dokumen serta seberapa unik kata tersebut dibandingkan dengan dokumen lain.

Berikut merupakan hasil hasil evaluasi terkait kinerja model yang dilatih menggunakan data hasil representasi teks TF-IDF.

Tabel 1. Classification Report (TF-IDF)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
|  |  |  |  |  |
| 0 | 0.89 | 0.89 | 0.89 | 10829 |
| 1 | 0.86 | 0.88 | 0.87 | 7790 |
| 2 | 0.81 | 0.78 | 0.80 | 6856 |
|  |  |  |  |  |
| Accuracy |  |  | 0.86 | 25475 |
| Macro AVG | 0.85 | 0.85 | 0.85 | 25475 |
| Weighted AVG | 0.86 | 0.86 | 0.86 | 25475 |

Berikut juga merupakan visualisasi dalam diagram batang yang menampilkan distribusi data berdasarkan kelas masing-masing.

A graph of different colored bars

Description automatically generated

Gambar 2. Visualisasi Distribusi Kelas Data (TF-IDF)

Adapun rincian distribusi kelas data ditampilkan pada tabel berikut:

Tabel 2. Rincian Distribusi Data (TF-IDF)

|  |  |
| --- | --- |
| Kelas | Jumlah |
| Politic | 53447 |
| Science | 40348 |
| Technology | 33580 |

**3.2. Klasifikasi Menggunakan Teknik Bag of Words**

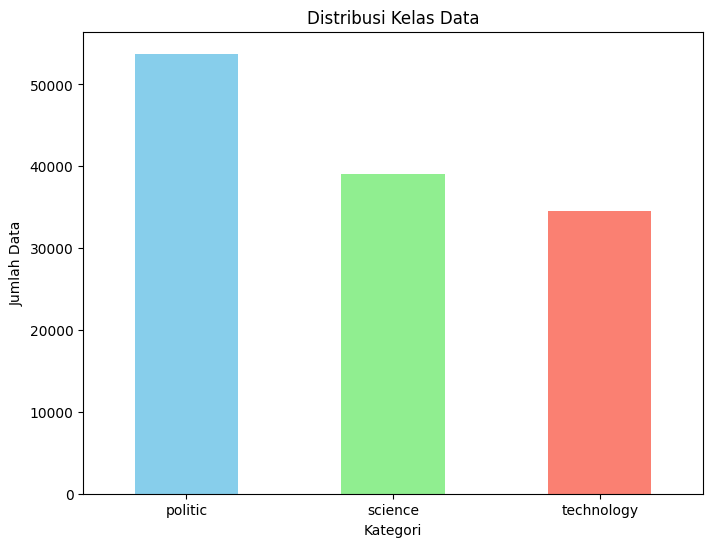
Bag of Words digunakan sebagai teknik representasi teks digunakan sebagai teknik representasi teks untuk mengubah data teks menjadi bentuk numerik sehingga dapat membantu model Logistic Regression dalam proses klasifikasi. Bag of Words berfungsi untuk memberikan bobot pada setiap kata berdasarkan frekuensi Teknik ini tidak memperhatikan urutan kata, hanya frekuensi atau keberadaannya.

Berikut merupakan hasil hasil evaluasi terkait kinerja model yang dilatih menggunakan data hasil representasi teks TF-IDF.

Tabel 3. Classification Report (Bag of Words)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
|  |  |  |  |  |
| 0 | 0.88 | 0.89 | 0.88 | 10829 |
| 1 | 0.86 | 0.85 | 0.85 | 7790 |
| 2 | 0.78 | 0.78 | 0.78 | 6856 |
|  |  |  |  |  |
| Accuracy |  |  | 0.85 | 25475 |
| Macro AVG | 0.84 | 0.84 | 0.84 | 25475 |
| Weighted AVG | 0.85 | 0.85 | 0.85 | 25475 |

Berikut juga merupakan visualisasi dalam diagram batang yang menampilkan distribusi data berdasarkan kelas masing-masing.



Gambar 3. Visualisasi Distribusi Kelas Data (Bag of Words)

Adapun rincian distribusi kelas data ditampilkan pada tabel berikut:

Tabel 4. Rincian Distribusi Data (Bag of Words)

|  |  |
| --- | --- |
| Kelas | Jumlah |
| Politic | 53588 |
| Science | 39316 |
| Technology | 34471 |

**3.3. Perbandingan Teknik TF-IDF dan Bag of Words**

Analisis terakhir yang dilakukan merupakan perbandingan teknik TF-IDF dan Bag of Words dalam membantu model Logistic Regression mengklasifikasikan kategori berita, mulai dari akurasi hingga distribusi data. Berikut merupakan perbandingan teknik TF-IDF dan Bag of Words dalam hal akurasi dan distribusi data.

Tabel 5. Perbandingan Akurasi

|  |  |
| --- | --- |
| Teknik Representasi Teks | Akurasi |
| TF-IDF | 0.86 |
| Bag of Words | 0.85 |

Berikut merupakan perbandingan rincian distribusi data menggunakan teknik TF-IDF dan Bag of Words:

Tabel 6. Perbandingan Rincian Distribusi Data

|  |  |  |  |
| --- | --- | --- | --- |
| Teknik Representasi Teks | Jumlah | | |
| Politic | Science | Technology |
| TF-IDF | 53588 | 39316 | 34471 |
| Bag of Words | 53447 | 40348 | 33580 |

**KESIMPULAN**

Berdasarkan perbandingan akurasi model Logistic Regression yang dilatih menggunakan data hasil vektorisasi TF-IDF dan Bag of Words, teknik representasi TF-IDF menjadi teknik terbaik dalam penelitian ini dengan akurasi mencapai 0.86.