**LAPORAN PROJECT AKHIR TEXT WEB MINING KELOMPOK 5**



**Disusun oleh**

Bryan Ozora C14220070

Darren Christopher C14220119

Benedict Janvier C1420176

Nicholas Daniel C14220207

**Program Studi Informatika**

**Universitas Kristen Petra**

**2025**

# Augmentasi Data (LLM)

Sebelum melakukan augmentasi data, perlu dilakukan labeling data yang dimana setiap anggota kelompok dan beberapa rekan sekelas melakukan data labelling guna membuat data asli yang akan digunakkan untuk training

Setelah melakukan labeling data, dilakukan augmentasi data dengan memanfaatkan LLM untuk membuat data baru berdasarkan data yang sudah ada. Hal ini dilakukan dengan mengubah bahasa dari salah satu review dari bahasa indonesia > bahasa inggris > bahasa random > bahasa inggris > bahasa indonesia. Bahasa random meliputi Spanish, Russian, Arabic, French, Azerbaijani, Chinese, Afrikaans, Afrikaans, Catalan, Danish, German, Finnish, Galician, Greek, Hungarian, Italian, Japanese.

Model LLM yang digunakkan adalah MarianMTModel yang merupakan model milik transformer dan hugging face yang merupakan model LLM yang dikembangkan untuk melakukan translasi. MarianMTModel ini menggunakkan berbagai model untuk melakukan translasi dari bahasa satu ke bahasa lainnya. Berikut adalah list model yang dapat digunakkan <https://huggingface.co/models?sort=trending&search=helsinki-NLP>

# Pembuatan Data Input

Data input didapatkan dari dataset yang telah diberikan kemudian dibagi menjadi 80:20 (16.000 : 4.000). Hal ini dilakukan untuk membagi data menjadi data training dan data testing, dimana 80 merupakan data training dan 20 testing.

Data training yang awalnya 80 (16.000) dibandingkan 20 (4.000) yang merupakan data testing, dilakukan augmentasi data. Setelah dilakukan augmentasi, data yang aslinya 16.000 berubah menjadi sekitar 31.000, namun data testing yang 4.000 tidak di-apa apakan dan ratio berubah menjadi sekitar 89.6:10.4.

# Suicide Ideation Prediction (BI-LSTM, BI-RNN, Transformer)

Suicide ideation prediction dilakukan dengan menggunakkan 3 metode yaitu dengan Bidirectional LSTM, Bidirectional RNN dan Transformer dengan tujuan untuk membandingkan kemampuan dan tingkat akurasi.

Suicide ideation dilakukan melalui beberapa step

1. **Tokenisasi dan Embedding**

Tokenisasi dan Embedding dilakukan pada data input guna untuk menghasilkan embedding dari review menggunakkan DistilBERT

1. **Data Splitting**

Data dipecah menjadi training dan testing dengan rasio 89,6:10,4 setelah augmentasi. Dimana X, merupakan kumpulan data dengan rasio 89.6 dan data Y adalah data dengan rasio 10.4

1. **Arsitektur Model (Perbedaan antar program model hanya disini)**

Model menggunakan lapisan Bi-LSTM / Bi-RNN / Transformer untuk menangkap hubungan sekuensial pada data, dengan dropout untuk menghindari overfitting.

1. **Training**

Model di-train dengan menggunakkan dataset X melalui 100 epoch untuk hasil maksimal

1. **Testing**

Setelah model di train, testing dilakukan pada data Y untuk di cek keakuratannya. Kemudian program akan menampilkan confusion matrix, akurasi, recall dan lain lain.

# Hasil

Setelah menjalankan semua program dengan settingan yang sama (perbedaan hanya pada layer model), kami mendapatkan hasil sebagai berikut

* **Bi-LSTM**

|  | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| 0 | 0.92 | 0.95 | 0.94 | 1440 |
| 1 | 0.93 | 0.90 | 0.91 | 1283 |
| 2 | 0.90 | 0.89 | 0.90 | 1323 |

Accuracy: 0.92

Confusion Matrix

|  | Predicted | | |
| --- | --- | --- | --- |
| Actual | 1374 | 24 | 42 |
| 42 | 1150 | 91 |
| 72 | 69 | 1182 |

* **Bi-RNN**

|  | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| 0 | 0.78 | 0.89 | 0.83 | 1440 |
| 1 | 0.74 | 0.85 | 0.79 | 1283 |
| 2 | 0.82 | 0.57 | 0.67 | 1323 |

Accuracy: 0.77

Confusion Matrix

|  | Predicted | | |
| --- | --- | --- | --- |
| Actual | 1282 | 95 | 63 |
| 83 | 1091 | 109 |
| 285 | 280 | 758 |

* **Transformer**

|  | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| 0 | 0.88 | 0.95 | 0.91 | 1440 |
| 1 | 0.88 | 0.89 | 0.88 | 1283 |
| 2 | 0.89 | 0.79 | 0.84 | 1323 |

Accuracy: 0.88

Confusion Matrix

|  | Predicted | | |
| --- | --- | --- | --- |
| Actual | 1362 | 37 | 41 |
| 49 | 1138 | 89 |
| 144 | 128 | 1051 |

# Kesimpulan

Berdasarkan hasil pengujian, performa tiga arsitektur model (Bi-LSTM, Bi-RNN, dan Transformer) untuk prediksi suicide ideation menunjukkan perbedaan yang signifikan, terutama dalam hal akurasi, presisi, recall, dan F1-Score.

1. **Bi-LSTM**
   * Bi-LSTM menunjukkan performa terbaik dengan akurasi tertinggi yaitu 92%.
   * Semua metrik (precision, recall, dan F1-Score) untuk ketiga kelas secara konsisten tinggi, terutama untuk kelas mayoritas (support terbesar).
   * Confusion matrix menunjukkan prediksi yang relatif stabil dengan jumlah kesalahan prediksi yang lebih rendah dibandingkan model lain.
2. **Bi-RNN**
   * Bi-RNN memiliki performa terendah dengan akurasi hanya 77%.
   * Kelas 2 menunjukkan hasil yang kurang memadai dengan F1-Score hanya 0.67, yang disebabkan oleh recall rendah (0.57).
   * Model ini cenderung kesulitan menangkap hubungan sekuensial yang kompleks dibandingkan dua model lainnya, seperti terlihat dari distribusi kesalahan pada confusion matrix.
3. **Transformer**
   * Transformer berada di posisi kedua dengan akurasi 88%.
   * Performanya lebih baik daripada Bi-RNN dalam semua metrik, namun masih sedikit tertinggal dari Bi-LSTM.
   * Kelemahan utama Transformer terlihat pada kelas 2, yang memiliki recall lebih rendah (0.79) dibanding kelas lain, meskipun precision untuk semua kelas cukup baik.

Kesimpulan Utama  
Bi-LSTM adalah arsitektur yang paling efektif untuk tugas ini, karena mampu menangkap hubungan sekuensial dengan baik sambil menjaga stabilitas performa antar kelas. Sementara itu, Bi-RNN kurang mampu menangkap kompleksitas data, sehingga menghasilkan performa terendah. Transformer menawarkan performa yang kompetitif, tetapi belum dapat mengungguli Bi-LSTM dalam konteks tugas prediksi suicide ideation.