Analisis Demografi pada Twitter Menggunakan Metode *Predictive Analytics*

**Kelompok Medsosan:**

Daniel Syahputra Purba - 1906398793

Fadli Aulawi Al Ghiffari - 1906285623

Muhammad Ichsanul Amal - 1906353454

**I. Latar Belakang**

Twitter adalah salah satu media sosial yang cukup banyak digunakan hampir oleh seluruh kalangan. Setiap user yang menggunakan Twitter bisa menyimpan beberapa informasi berupa atribut demografi di masing-masing profilnya. Meskipun begitu tidak semua user melengkapi informasi-informasi tersebut, bahkan beberapa user mengisinya dengan data palsu. Padahal jika ditinjau lebih jauh, data demografi ini bisa digunakan untuk beberapa keperluan seperti perancangan marketing plan, atau analisis dukungan masyarakat terkait kontestasi politik yang sedang berjalan.

Mengatasi permasalahan tersebut, kami ingin melakukan sebuah penelitian dengan suatu pendekatan prediktif. Atribut demografi pada tiap akun di Twitter akan diprediksi menggunakan data lainnya di akun tersebut seperti tweet, atau atribut demografi lainnya yang tersedia. Pada penelitian ini akan terkhusus pada prediksi dua atribut, yaitu *gender* dan *occupation* (pekerjaan). Prediksi akan dilakukan dengan berbagai cara yang memiliki konsep dasar *supervised learning* dan *rule-based learning*.

**II. Metodologi dan Analisis Metode**

Dalam melakukan penelitian, peneliti melaksanakan serangkaian tahapan yang terbagi menjadi beberapa bagian, yaitu:

## **a. Riset**

Penelitian dimulai dengan melakukan observasi dan mempelajari riset yang telah ada sebelumnya terkait analisis demografi. Hal ini dilakukan untuk memperkaya pengetahuan peneliti, sehingga dapat menghemat waktu implementasi saat perancangan model. Selain menggunakan pengetahuan terkait analisis sosial media, peneliti juga mempelajari pengetahuan yang berkaitan dengan *Natural Language Processing* dan *Text Mining*. Hal ini dikarenakan pada penelitian ini data yang digunakan merupakan data *unstructured* berupa teks, sehingga pengetahuan terkait pengolahan teks sangat diperlukan dalam penelitian ini.

Selain melakukan observasi dan penelusuran terkait hal-hal terkait konsep pengolah teks dan analisis media sosial, peneliti juga melakukan penelusuran terkait hal teknis seperti penggunaan *tools* dan *libraries* yang dapat membantu saat proses pengolahan data tersebut. Adapun *tools* atau *library* yang peneliti rasa dapat membantu penelitian kedepannya seperti Sastrawi, *pre-trained model* seperti IndoBERT atau Word2Vec yang telah dilatih dengan jutaan teks dalam bahasa Indonesia, dll.

## **b. Eksplorasi Data**

Setelah peneliti melakukan riset terkait permasalahan, peneliti melakukan eksplorasi dan observasi terhadap data yang diberikan. Hal ini bertujuan untuk memberikan pemahaman awal terkait data seperti adanya *missing-values*, apakah terdapat kasus *imbalanced class* pada data kepada peneliti. Dengan adanya pengetahuan tersebut, hal ini akan sangat membantu peneliti nantinya dalam merancang model yang lebih *robust.*

## **c. Pembersihan Data dan Pra-proses Data**

Berdasarkan informasi yang didapatkan pada tahapan sebelumnya, peneliti melakukan pembersihan data dikarenakan memang terdapat banyak *missing-values* seperti terdapat beberapa profil yang tidak berisikan bidang pekerjaan yang digeluti. Dikarenakan penelitian tidak memiliki pengetahuan terkait profil-profil tersebut, maka peneliti memutuskan hanya menggunakan data yang berisikan informasi lengkap pada penelitian ini.

Data yang telah melalui tahap pembersihan selanjutnya akan melalui tahap pra-pemrosesan. Tahapan ini bertujuan untuk menghasilkan teks yang lebih bersih dan hanya mengandung hal-hal yang bermanfaat untuk proses analisis saja. Adapun proses yang dilakukan pada tahapan ini meliputi *text replacement* menggunakan *regular expression*, penghapusan *stopwords,* mengganti kata singkatan atau slang, dan juga *stemming*.

## **d. Pemodelan**

Data yang telah melalui tahapan pembersihan dan pra-pemrosesan akan digunakan untuk proses pemodelan. Namun, sebelum melakukan pemodelan, data tersebut harus diubah terlebih dahulu ke dalam representasi numerik / vektor agar dapat digunakan dalam pemodelan. Pada penelitian ini, peneliti menggunakan beberapa pendekatan dalam membentuk representasi teks, yaitu pendekatan *Bag of Words* seperti *Term-Frequency* dan *Term-Frequency and Inverse Document Frequency* dan juga pendekatan *Word2Vec* dan *Transformer.*

Setelah data telah direpresentasikan ke dalam bentuk numerik, data dapat digunakan untuk pemodelan. Pada penelitian ini, peneliti menggunakan pendekatan *supervised learning* dan *rule-based*. Pada pendekatan *rule based* akan dilakukan prediksi gender berdasarkan nama seseorang. Pemodelan akan menggunakan data yang telah dibagi dengan proporsi tertentu menjadi data *training* dan *validation*.

## **e. Evaluasi Model**

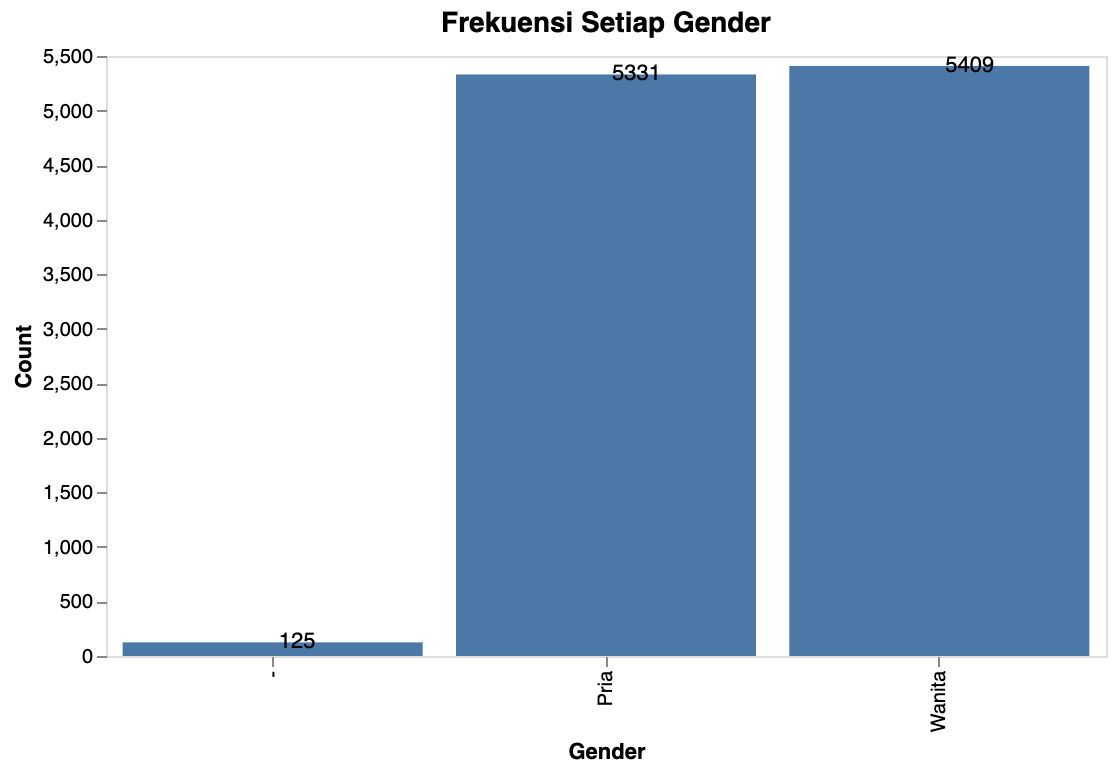
Setelah dilakukan pemodelan, akan dilakukan evaluasi model menggunakan data *validation* untuk menentukan apakah model sudah cukup baik atau tidak. Adapun metrik yang kami gunakan dalam proses evaluasi meliputi *F1-Score, Recall, Precision,* dan *Accuracy.*

**III. Penjelasan *Exploratory Data Analysis***

Data yang digunakan pada penelitian ini berasal dari twitter yang terbagi dua jenis yaitu *structured* dan *unstructured.* Untuk data yang bertipe *structured*, data berisikan informasi terkait profil seseorang seperti *gender, description, followers\_count, occupation, following\_count,* dll.

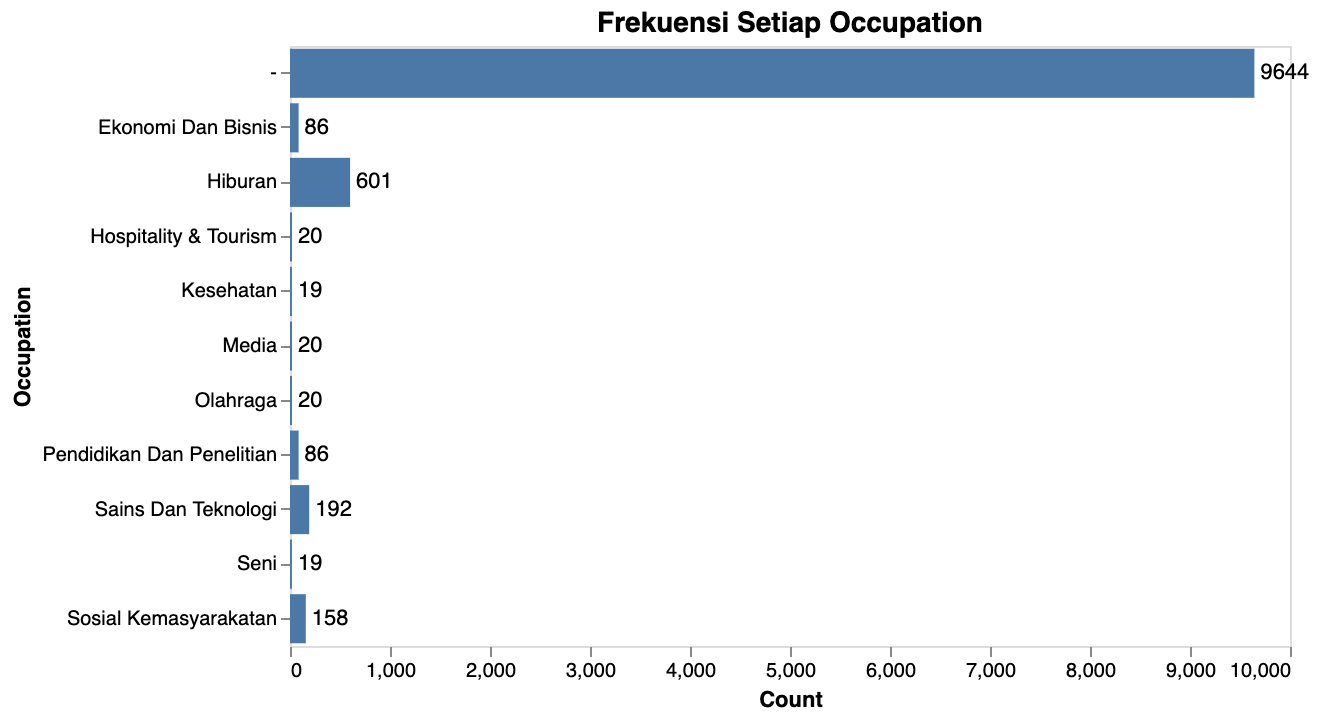
## **a. Gender**

Berikut ini adalah persebaran dari setiap gender pada dataset yang dimiliki. Dari informasi dibawah, dapat diketahui bahwa terdapat 125 profil yang *gender* nya belum diketahui. Selain itu dapat dilihat juga bahwa frekuensi untuk *gender* “Pria” dan “Wanita” cukup seimbang, sehingga kita bisa memastikan saat melakukan prediksi gender tidak terdapat kasus *imbalanced classification.*



## **b. Occupational**

Selain melakukan prediksi *gender*, peneliti juga melakukan prediksi *occupational* seseorang berdasarkan profil yang diberikan. Pada penelitian ini terdapat sepuluh jenis pekerjaan dan peneliti akan mencoba memprediksi pekerjaan seseorang baik itu berdasarkan *tweet*, *description*, dll. Berikut ini adalah grafik frekuensi setiap pekerjaan pada dataset:



Berdasarkan grafik diatas, dapat dilihat bahwa terdapat *imbalanced* pada dataset yang dimana terdapat pekerjaan dengan jumlah yang sangat sedikit dibandingkan yang lain. Oleh sebab itu, hal ini akan menjadi pertimbangan nantinya ketika melakukan pemodelan. Selain itu, pada dataset banyak terdapat profil yang *occupation* nya belum diketahui, sehingga pada penelitian ini hanya akan digunakan profil yang *occupation* nya telah diketahui saja.

## **c. Eksplorasi lain**

Rata-rata tiap fitur

| **Gender** | **followers\_count** | **following\_count** | **listed\_count** | **tweet\_count** |
| --- | --- | --- | --- | --- |
| Wanita | 38.323 | 772 | 709 | 15.735 |
| Pria | 70.305 | 629 | 55 | 11.292 |

Jumlah maksimal tiap fitur pada tiap gender

| **Gender** | **followers\_count** | **following\_count** | **listed\_count** | **tweet\_count** |
| --- | --- | --- | --- | --- |
| Wanita | 1.7875.098 | 301.883 | 3.095.544 | 15.735 |
| Pria | 7.936.5018 | 196.047 | 88.198 | 11.292 |

Jumlah akun *verified* pada tiap gender

| **Gender** | **Jumlah** |
| --- | --- |
| Wanita | 86 |
| Pria | 163 |

**IV. Cara Kerja dan Tahapan Mengolah Data**

Pada penelitian ini dilakukan beberapa percobaan. Prediksi *gender* dan *occupation* dilakukan dengan menggunakan dataset *unstructured* *tweet*, dataset *structured profile*, dan gabungan dari keduanya. Kedua dataset yaitu *tweet* dan profil akun digabungkan menjadi satu dengan *join* kolom “username”. Setelah itu, dilakukan *data preprocessing* dengan langkah sebagai berikut

## **a. Membersihkan kolom *tweet* dan *description***

Pembersihan kolom *tweet* dan *description* mencakup menghilangkan digit, transformasi semua huruf ke *lowercase*, menghapus URL, *username*, rt, sep, *white space*, dan pembersihan lainnya.

## **b. Transformasi bahasa dan menghapus *stopwords***

Transformasi bahasa mencakup transformasi kalimat yang mengandung singkatan menjadi kalimat yang baku, kalimat dengan ragam bahasa tidak baku menjadi kalimat dengan ragam bahasa baku, dan penghapusan *stopwords*. Kamus singkatan didapatkan dari *kamus\_singkatan.csv*, kamus ragam bahasa tidak baku didapatkan dari *colloquial-indonesian-lexicon.csv,* dan *stopwords* yang digunakan berasal dari *stopwords.csv*. Kamus-kamus yang kami sebutkan sebelumnya, kami dapatkan dari sumber luar yaitu Github.

## **c. *Stemming***

Pada awalnya, peneliti akan menjalankan langkah ini. Namun, proses *stemming* memerlukan waktu yang cukup lama sehingga peneliti melewati langkah ini. Setelah dilakukan penelitian, ternyata tanpa dilakukan langkah ini pun hasilnya sudah cukup memuaskan.

## **d. *Feature Extraction***

Agar teks yang dimiliki dapat digunakan dalam pemodelan, teks tersebut harus diubah ke dalam representasi numerik. Ada beberapa representasi dan teknik yang dapat digunakan sebagai berikut:

1. CountVectorizer atau Term Frequency (TF)

CountVectorizer digunakan untuk mengubah teks yang diberikan menjadi vektor berdasarkan frekuensi (jumlah) setiap kata yang muncul di seluruh teks.

1. TfidfVectorizer atau Term Frequency–Inverse Document Frequency (TF-IDF)

TF IDF atau pembobotan kata merupakan skema yang digunakan untuk menghitung bobot setiap kata yang paling umum digunakan [3]. Pada TF-IDF dibagi menjadi dua bagian yaitu :

* 1. TF (Term Frequency) merupakan frekuensi kemunculan sebuah *term* dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu *term* dalam dokumen, maka semakin besar bobotnya.
  2. IDF (Inverse Document Frequency) merupakan sebuah perhitungan dari bagaimana *term* didistribusikan secara pada koleksi dokumen yang bersangkutan. Semakin sedikit jumlah dokumen yang mengandung *term* yang dimaksud, maka nilai IDF semakin besar.

1. Word2Vec (Word Embeddings)

Word Embeddings adalah proses konversi kata ke dalam bentuk vektor. Setiap kata adalah vektor yang merepresentasikan sebuah titik pada ruang dengan dimensi tertentu. Dengan Word Embeddings, kata-kata yang memiliki properti tertentu, atau memiliki *semantic meaning* yang sama berada tidak jauh satu sama lain pada ruang tersebut [2]. Sedangkan Word2Vec merupakan salah satu adalah model *shallow neural network* yang merubah representasi kata yang merupakan kombinasi dari karakter *alphanumeric* menjadi vektor. Representasi vektor tersebut memiliki properti *relationship* terhadap kata-kata yang berkaitan melalui proses *training*.

1. Transformer dengan IndoBERT Large

IndoNLU mencakup 12 *task* yang dibagi lagi menjadi 4 kategori, yaitu klasifikasi kalimat tunggal, penandaan urutan kalimat tunggal, klasifikasi pasangan kalimat, dan pelabelan urutan pasangan kalimat [1]. Peneliti mencoba menggunakan *transformer* ini karena disebutkan bahwa *transformer* ini memiliki kemampuan yang cukup baik dalam masalah *Indonesian Natural Language Understanding*. Kode yang digunakan untuk percobaan ini berasal dari *library* yang telah dibuat oleh Indo Benchmark [4].

## **e. Pemodelan**

Pada tahapan *training* dilakukan *train test split* terlebih dahulu dengan ukuran data untuk *validation* sebesar 30% untuk kebutuhan *tuning* dari model. Setelah itu akan dilakukan pemodelan dengan berbagai teknik seperti:

1. Pemodelan *gender* dan *occupation* menggunakan data profil yaitu *description*
2. Pemodelan *gender* dan *occupation* menggunakan data *tweet*
3. Pemodelan *gender* menggunakan nama dengan tambahan data *training* dari sumber lain.
4. Pemodelan *gender* dan *occupation* dengan gabungan data profil dan *tweet*.

Pada pemodelan, peneliti menggunakan pendekatan *supervised learning* dan *rule-based*. Adapun pemodelan menggunakan beberapa algoritma seperti *Support Vector Machine, Random Forest, KNN, Adaboost,* hingga model *deep learning seperti* LSTM (*Long Short Term Memory*)*.*

Untuk setiap teknik pemodelan yang telah disebutkan peneliti sebelumnya, akan dilakukan percobaan dengan berbagai tipe representasi baik itu TF, TF-IDF, representasi dengan Word2Vec atau IndoBERT.

Selain itu, peneliti juga memanfaatkan sumber dari luar khususnya pada saat pemodelan gender menggunakan metode *rule based* berdasarkan nama. Peneliti menggunakan bantuan dataset *indonesian-gender-name-dataset* yang telah tersedia pada Github [5]. Tambahan dataset ini akan digabungkan dengan dataset nama dari data *profile* untuk dijadikan data *training* dan *validation.*

**V. Hasil Implementasi**

## **a. Prediksi Gender**

Berikut ini adalah perbandingan nilai evaluasi pada data *validation* dari setiap metode yang telah dilakukan oleh peneliti dalam memprediksi *gender*:

| **Data**  **Training** | **Model** | **Accuracy** | **Recall** | **Precision** | **F1-Score** |
| --- | --- | --- | --- | --- | --- |
| Tweet + Profile + Transformer | Support Vector Machine | 0.6 | 0.6 | 0.6 | 0.6 |
| Random Forest | 0.6 | 0.6 | 0.6 | 0.6 |
| Ada Boost | 0.58 | 0.58 | 0.58 | 0.58 |
| Multi Layer Perceptron | 0.59 | 0.60 | 0.61 | 0.58 |
| K-Nearest Neighbors | 0.58 | 0.58 | 0.58 | 0.58 |
| Tweet + TF-IDF [BEST] | Support Vector Machine | 0.63 | 0.63 | 0.63 | 0.63 |
| Random Forest | 0.65 | 0.65 | 0.65 | 0.65 |
| AdaBoost | 0.62 | 0.62 | 0.62 | 0.62 |
| Profile Name + TF | Support Vector Machine | 0.72 | 0.72 | 0.74 | 0.71 |
| Random Forest | 0.72 | 0.72 | 0.75 | 0.71 |
| Profile Name + Description + Transformer | Support Vector Machine | 0.55 | 0.55 | 0.55 | 0.55 |
| Random Forest | 0.55 | 0.55 | 0.55 | 0.54 |

Dari hasil diatas, peneliti mendapatkan bahwa penggunaan data *tweet* yang dikombinasikan dengan representasi TF-IDF mendapatkan hasil terbaik ketika dihadapkan data test di *grader.*

## **b. Prediksi Occupation**

Berikut ini perbandingan nilai evaluasi dari setiap metode yang telah dilakukan untuk memprediksi *occupation*:

| **Data Training** | **Model** | **F1-Score** |
| --- | --- | --- |
| Data *Profile* Description (Sebagian *class*) + Transformer Large [BEST] | Support Vector Machine | 0.498 |
| Random Forest | 0.447 |
| Data Profil (Semua *class*) | Support Vector Machine | 0.06 |
| Random Forest | 0.19 |

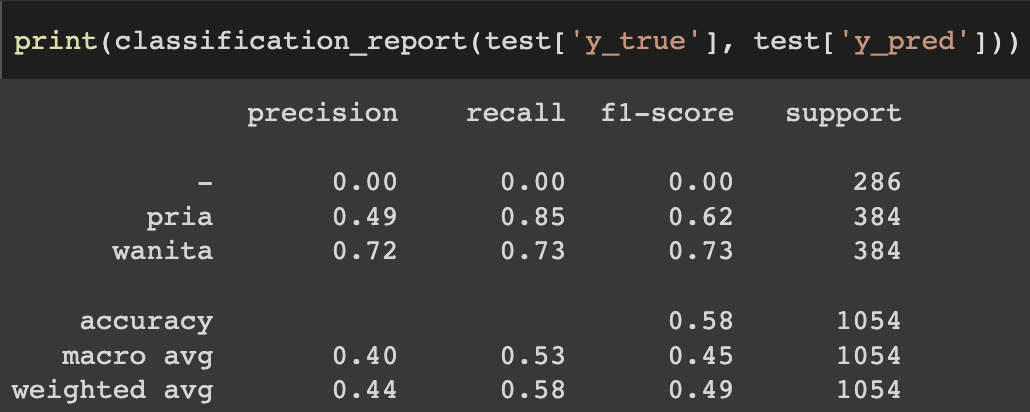
Dari hasil diatas, peneliti mendapatkan bahwa penggunaan data profil *description* dikombinasikan dengan *transformer model* dalam membentuk representasi teks, mendapatkan hasil terbaik pada data test. Pada penelitian ini, peneliti menggunakan dua jenis ukuran *transformer* dari IndoBERT yaitu yang *small* dan large dan didapatkan bahwa hasil dengan menggunakan *pre-trained* model yang berukuran *large* memberikan hasil terbaik.

**VI. Analisis Error**

Setelah dilakukan pengetesan pada data *testing,* peneliti kemudian melakukan *error analysis* untuk mengevaluasi apakah terdapat alasan mengapa performa model masih belum terlalu baik.

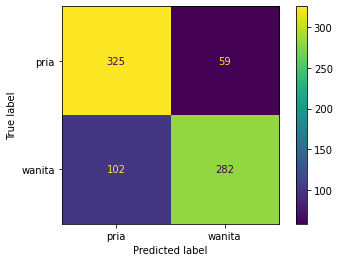
## **a. Analisis Gender**

Berikut ini adalah hasil apabila membandingkan hasil prediksi dengan *gold standard* pada prediksi *gender*.



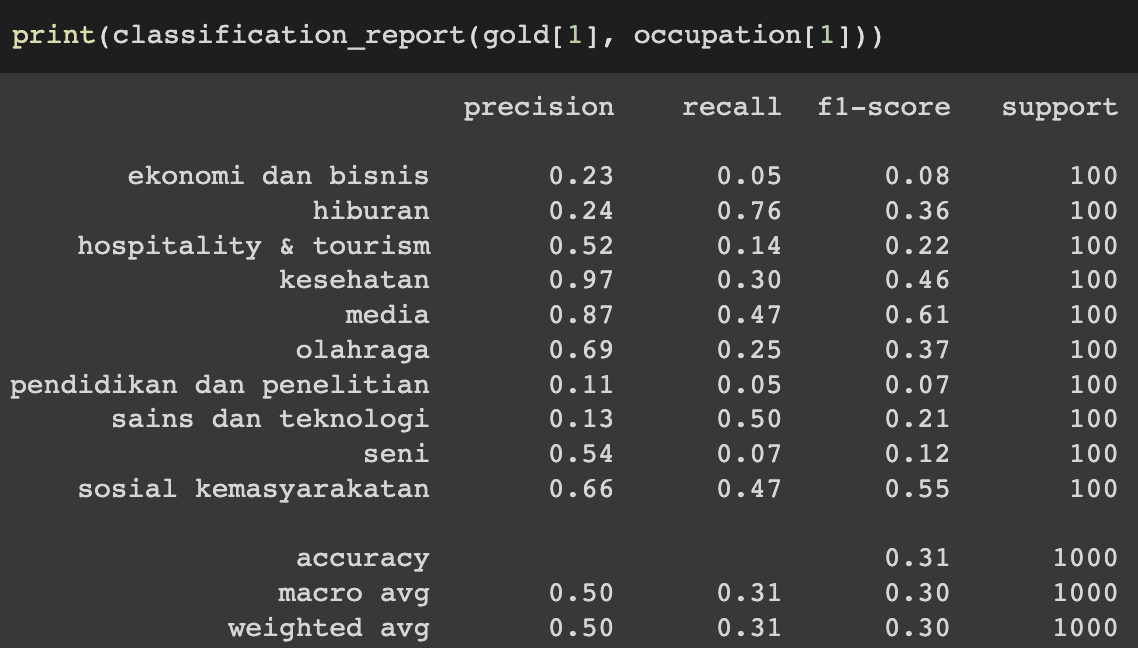
Dari hasil diatas, hasil yang didapatkan tidak sebaik pada grader karena peneliti masih mengikutsertakan label ‘-’ pada *gold standard*. Berikut ini adalah hasil evaluasi apabila yang dilihat label yang tidak ‘-’ pada *gold standard.*





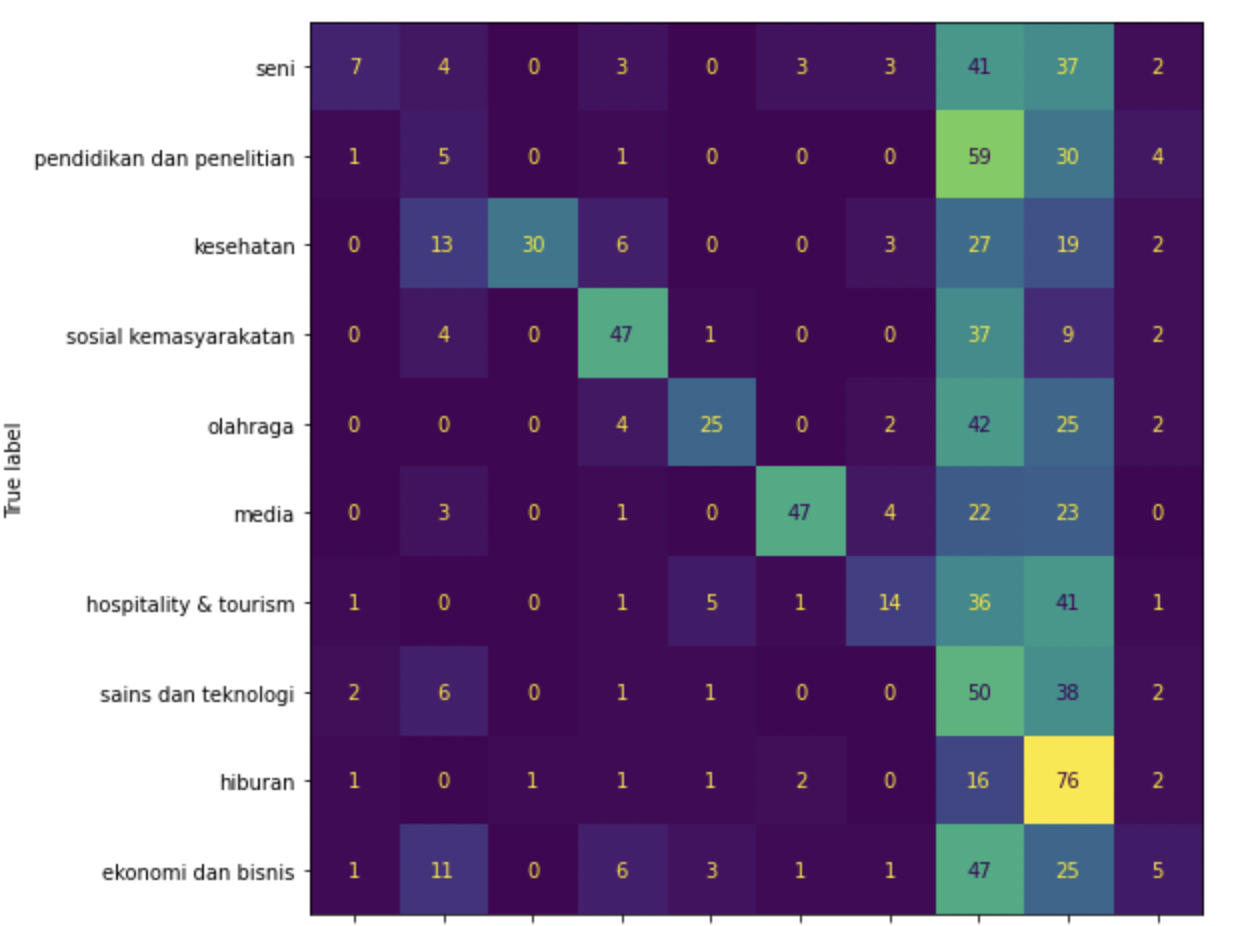
## **b. Analisis Occupation**

Berikut ini adalah hasil evaluasi dari hasil prediksi terbaik pada *occupational* dengan *gold standard*:



Berdasarkan hasil diatas, dapat dilihat bahwa terdapat *class* yang F1-Score nya bernilai nol. Hal ini dikarenakan peneliti hanya memanfaatkan sebagian class untuk *training* saat melakukan pemodelan. Apabila data yang digunakan saat *training* memiliki *class* yang lengkap, hasil yang didapatkan justru kurang baik karena semakin banyaknya nilai probabilitas yang terbagi.

Oleh karena itu, peneliti memutuskan hanya menggunakan data yang berisikan sebagian *class* tertentu, kemudian memaksimalkan prediksi pada *class* ini dibanding melakukan prediksi ke semua *class.* Berikut ini adalah *confusion matrix* dari perbandingan hasil prediksi dengan *gold standard*:



## **c. Kekurangan dalam proses implementasi**

Pada saat melakukan *text preprocessing,* peneliti melewatkan tahap *stemming* dikarenakan beban waktu komputasi. Hal ini mengakibatkan terlalu banyak kata unik pada teks bahkan setelah dilakukan *preprocessing* sehingga mengakibatkan matrix semakin *sparse* apabila menggunakan metode *text representation* berbasis *bag of words.* Selain itu, hal ini mengakibatkan banyaknya kata yang dianggap *out of vocabulary (OOV)* ketika melakukan *word embedding* dengan IndoBERT dan menghasilkan hasil *embedding* yang kurang baik.

Pada proses pemodelan untuk prediksi *occupational*, peneliti hanya memfokuskan sebagian *class* pada *training* nya. Hal ini mengakibatkan banyaknya *class* yang memiliki F1-Score bernilai nol. Meskipun hasil dari prediksi *occupational* menggunakan metode ini cukup baik, hal ini bisa menjadi *concern* kedepannya bagaimana merancang model yang lebih baik dengan mengikutsertakan setiap *class*.

**VII. Kesimpulan**

Dari percobaan yang kami adakan diperoleh hasil bahwa model yang terbaik untuk prediksi *gender* didapat dari model supervised (Random Forest) yang dilatih dengan data teks *tweet* dan direpresentasikan dengan TF-IDF, dan model terbaik untuk prediksi occupation didapat dari model supervised (Support Vector Machine) yang dilatih dengan data teks *description* pada profil dan direpresentasikan dengan Transformer Large. Model tersebut dipilih dari perbandingan skor evaluasi dengan model-model lainnya yang sudah dicoba sebelumnya.

Meskipun begitu, masih terdapat beberapa kekurangan dari pengerjaan yang kami lakukan. Salah satunya adalah kami belum mencoba seluruh kemungkinan model yang bisa digunakan pada proses prediksi label ini.. Diharapkan dengan banyaknya variasi model yang dieksplorasi bisa ditemukan model yang lebih baik performanya dalam memprediksi label *gender* dan *occupation.*

Terakhir kami menyadari belum adanya proses *balancing data*, padahal terdapat beberapa label yang perbandingannya cukup *imbalanced* antara data yang memiliki label tersebut dengan yang tidak. Ketidakimbangan data ini dapat ditangani dengan beberapa metode seperti *oversampling* dengan library SMOTE, *undersampling* dengan library Imblearn, atau penerapan *class-weighting* pada model-model prediksi yang dibuat.

**VIII. Referensi**

[1] Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X., Lim, Z., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. AACL.

[2] Word Embedding dengan Word2vec. (2022). Retrieved 25 April 2022, from

<https://mti.binus.ac.id/2020/11/17/word-embedding-dengan-word2vec/>

[3] Munawar, Y. R. S. Sistem Pendeteksi Berita Hoax di Media Sosial dengan Teknik Data Mining Scikit Learn. Retrieved 25 April 2022, from <https://digilib.esaunggul.ac.id/public/UEU-Journal-20004-11_1244.pdf>

[4] indobenchmark/indobert-large-p1· Hugging Face. (2022). Retrieved 25 April 2022, from <https://huggingface.co/indobenchmark/indobert-large-p1>

[5] indonesian-name-gender-dataset/name\_gender.csv at master· vck/indonesian-name-gender-dataset. (2022). Retrieved 25 April 2022, from <https://github.com/vck/indonesian-name-gender-dataset/blob/master/name_gender.csv>