Author Profiling Data Twitter

Mas Raihan, Shella, Ahmad Fauzi Wibowo, Gilang, Marisha Salsabila

Abstract— *Twitter* merupakan salah satu media sosial yang populer digunakan oleh masyarakat global. Masifnya data yang dimiliki oleh twitter menjadikannya sebagai sumber data yang dapat digunakan oleh peneliti maupun analis untuk memperoleh informasi di dalamnya. Data yang berupa *textual* tersebut dapat diolah menggunakan Natural Language Processings hingga dapat diperoleh informasi yang dibutuhkan salah satunya dalam melakuakan *Author Profiling*. Tujuan dari paper ini adalah untuk mendapatkan model training data tweet yang baik sehingga mampu memperoleh akurasi yang baik dalam melakukan *author profiling*. Dilakukan preprocessing data dengan melakukan HTML encoding, melakukan regex untuk fitur mention, hastag, maupun link. Dilakukan proses *feature selection* menggunakan K-Best Selection Feature dengan chi_square. Digunakan beberapa algoritma untuk menerapkan klasifikasi yang dilakukan diantaranya : Decission Tree, MLP, SVM dan RNN. Diperoleh hasil akurasi terbaik menggunakan algoritma SVM dengan akurasi sebesar 0.7201

Keyword : Twitter, Natural Language Processing, AuthorProfiling

I. PENDAHULUAN

A. Latar Belakang

Media sosial merupakan media yang dapat menghubungkan satu individu dengan individu lainnya untuk dapat bersosialsasi, bertukar informasi hingga menyampaikan pendapat secara online. Media sosial sangat populer digunakan mengingat pesatnya laju teknologi informasi yang berkembang saat ini. Melalui sosial media, masyarakat memperoleh kebebasannya untuk dapat saling berkomunikasi hingga mengemukakan pendapat, sehingga melalui media sosial diperoleh beraneka macam informasi yang didapatkan. Twitter merupakan salah satu media sosial yang populer digunakan masyarakat global dengan Amerika Serikat sebagai pengguna aktif twitter terbanyak kedua di dunia dengan total 240 juta user[1]. Setiap hari, jam bahkan detik jumlah kicauan pada twitter selalu bertambah. Oleh karena banyaknya jumlah data yang dimiliki, twitter banyak dijadikan sebagai sumber data bagi para peneliti maupun analis untuk melakukan pengolahan data sehingga dapat diperoleh informasi lebih di dalamnya.

Banyaknya data kicauan twitter yang mana berupa textual dapat diolah dengan menggunakan *Natural Language Processing* sehingga dapat diperoleh informasi yang dapat diaplikasikan ke berbagai macam aspek baik dalam dunia industri, strategi pemasaran, hingga penelitian di bidang personalisasi dalam lingkup psikologi maupun sosiologi[2]. Salah satu hal yang dapat dilakukan dengan menggunakan *Natural Language Processing* adalah untuk melakukan *author profiling* untuk membedakan mana user pria ataupun wanita.

Besarnya jumlah data kicauan pada twitter akan mengakibatkan besarnya pula biaya komputasi yang harus

dilakukan dengan proses *Natural Language Processing.* Pada paper ini akan dilakukan pengambilan data twitter yang akan dijadikan sebagai *training data* untuk dijadikan model pengklasifikasian dan *testing data* untuk menguji keakuratan model pelatihan data yang telah dimiliki. Oleh karena itu dibutuhkan model *training data* yang baik sehingga dapat dilakukan proses pemodelan yang efisien dan juga menghasilkan akurasi yang baik untuk melakukan *author profiling*.

B. Tujuan dan Manfaat

Berdasarkan permasalahan yang telah dikemukakan sebelumnya, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Memperoleh model data pelatihan yang baik untuk melakukan *author profiling* sehingga didapatkan akurasi yang baik.
2. Mengetahui algoritma klasifikasi yang cocok untuk digunakan dalam *author profiling.*

Diharapkan dengan penelitian ini tercapai beberapa manfaat sebagai berikut:

1. Membantu industri maupun pihak-pihak yang membutuhkan informasi mengenai *author profiling.*
2. Memberikan kontribusi pada ilmu pengetahuan dalam hal informasi model data yang baik dalam melakukan *author profiling.*

II. METODE

A. Perangkat Uji Coba

Dalam menyelesaikan permasalahan ini, digunakan paket perangkat lunak Anaconda dengan Spyder IDE sebagai editornya. Library yang digunakan meliputi numpy, pandas, scikit-learn, dan matplotlib. Bahasa pemrograman yang digunakan adalah Python 3.6.

Perangkat keras yang digunakan adalah sebuah Laptop Lenovo(™) ideapad(™) 710S dengan prosesor Intel(R) Core(™) i5-7260U, memori 8192 MB dan menggunakan sistem operasi Windows 10 Home 64-bit.

B. Dataset

Dataset yang digunakan merupakan data kicauan twitter, gender dan juga usia dari beberapa user twitter. Data kicauan yang digunakan menggunakan bahasa inggris. Data diperoleh dari sebuah kompetisi yang diselenggarakan oleh PAN : Author Profiling 2016 yang didownload dari https://pan.webis.de/clef16/pan16-web/author-profiling.html.

Dataset terdiri dari data train untuk dilakuakn pemodelan dan data test sebagai data untuk menguji model yang telah dibuat. Data train berbahasa inggris teridiri dari 5003 tweet yang berasal dari 440 user twitter. Data test terdiri dari blabla ratus tweet dari 78 user twitter untuk dilakukan pengujian terhadap model data train.

Pada data train hanya digunakan data kicauan tweet sebagai atribut dan gender sebagai target pengklasifikasian yang akan dilakukan.

C. Dasar Teori

(i) Natural Language Processing

Natural Language Processing (NLP) adalah area penelitian dan aplikasi yang mengeksplorasi bagaimana komputer dapat digunakan untuk memahami dan memanipulasi teks atau pidato bahasa natural untuk melakukan hal yang bermanfaat[3]. Aplikasi yang paling sering menggunakn NLP adalah sebagai berikut:

- *Information Retrieval* – menyediakan daftar dokumen yang berpotensi untuk relevan sebagai tanggapan atas permintaan pengguna.
- *Information Extraction* (IE) – berfokus pada pengenalan, penandaan, dan ekstraksi ke dalam representasi terstruktur, elemen kunci tertentu dari informasi, misalnya orang, perusahaan, lokasi, organisasi, dari koleksi teks yang banyak.

- *Question Answering* – menyediakan seorang pengguna hanya dengan teks jawaban itu sendiri atau bagan yang menyediakan jawaban.
- *Summarization* – mengurangi teks yang panjang menjadi pendek, namun tetap merepresentasikan dokumen aslinya.
- *Machine Translation* – aplikasi NLP yang paling tua. Berbagai tingkat NLP telah digunakan dalam sistem *Machine Translation*, mulai dari pendekatan '*word-based'* hingga aplikasi yang mencakup tingkat analisis yang lebih tinggi
- *Dialogue Systems – dialogue system* yang biasanya fokus pada aplikasi yang didefinisikan secara sempit (mis. kulkas atau sistem suara di rumah), saat ini sudah menggunakan tingkat bahasa *phonetic* dan *lexical.[4]*

Penelitian ini termasuk dalam pengaplikasian NLP berupa *Infromation Extraction* atau IE. Informasi yang akan diekstrak adalah jenis kelamin dari *author* akun Twitter, sehingga nantinya dapat membedakan antara *author* pria dengan *author* wanita.

(ii) Author Profiling

Author Profiling adalah sebuah kegiatan untuk memprediksi satu atau lebih ciri-ciri penulis dan profil penulis terdiri dari set yang dihasilkan dari satu atau lebih ciri-ciri yang diprediksi [3]. Tujuan dari *author profiling* adalah untuk memprediksi karakteristik spesifik dari *author* dengan cara menganalisis dokumen yang ditulisnya. Ada banyak karakteristik dari *author* yang dapat diprediksi, misalnya adalah umur dan jenis kelamin, yang mana paling banyak dibahas pada penelitian-penelitian terdahulu.

Author profiling dapat digunakan pada banyak bidang, tergantung ciri-ciri apa yang akan diprediksi. Dalam beberapa kasus, informasi mengenai ciri-ciri dari *author* bisa jadi sangat berharga. Misalnya pada bidang linguistik forensik, jika dapat diketahui profil linguistik dari sebuah pesan yang dicurigai dan diidentifikasi karakteristiknya hanya dengan menganalisis teks tersebut, maka akan sangat membantu dalam penelusuran tersangka pengirim pesan. Contoh lainnya pada bidang pemasaran, dengan menganalisis ulasan produk, nantinya akan dapat diprediksi produk apa yang disukai atau tidak disukai oleh orang-orang.

C. Algoritma

(i) Decision Tree

Pohon keputusan atau dikenal dengan Decision Tree adalah salah satu metode klasifikasi yang menggunakan representasi suatu struktur pohon yang berisi alternatif untuk pemecahan suatu masalah [5]. Decision tree terdiri dari tiga bagian yaitu : a) Root node. Node ini merupakan node yang terletak paling atas dari suatu pohon; b) Internal node. Node ini merupakan node percabangan, hanya terdapat satu input serta mempunyai minimal dua output; c) Leaf node. Node ini merupakan node akhir, hanya memiliki satu input dan tidak memiliki output.

Pohon dibangun dalam suatu metode rekursif topdown divide and conquer. Bentuk pemecahannya menggunakan

algoritma yang bergantung pada jenis atribut. Pada algoritma C-45 menggunakan gain ratio.

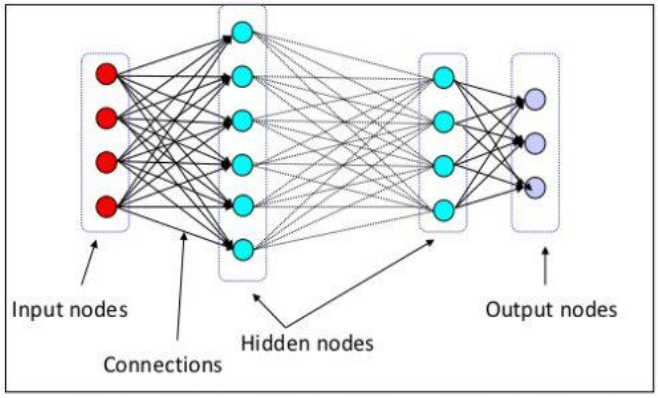
(ii) Random Forest

Random Forest merupakan pengembangan dari *Decision Tree* dengan melakukan pembuatan beberapa *Decision Tree,* dimana setiap tree telah dilakukan pelatihan menggunakan sampel individu dan setiap atribut dipecah pada tree yang dipilih antara atribut subset yang bersifat acak. Proses klasifikasi dilakukan pada masing-masing tree yang telah dibuat, kemudian dipilih berdasarkan vote/suara terbanyak pada kumpulan populasi tree [5].

(iii) MLP

Multi-layer Perceptron (MLP) adalah jaringan syaraf tiruan *feed-forward* yang terdiri dari sejumlah neuron yang dihubungkan oleh bobot-bobot penghubung.

Gambar II.1 Arsitektur MLP



Neuron-neuron tersebut disusun dalam lapisan-lapisan yang terdiri dari satu lapisan input *(input layer),* satu atau lebih lapisan tersembunyi *(hidden layer)*, dan satu lapisan output *(output layer).* Lapisan input menerima sinyal dari luar, kemudian melewatkannya ke lapisan tersembunyi pertama, yang akan diteruskan sehingga akhirnya mencapai lapisan output [6].

(iii) SVM

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikanpada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti margin hyperplane[] dan demikian juga dengan konsep-konsep pendukung yang lain.. Prinsip dasar SVM adalah linear classifier dan selanjutnya dikembangkan agar dapat bekerja pada problem non-linear dengan memasukkan konsep *kernel trick* pada ruang kerja berdimensi tinggi

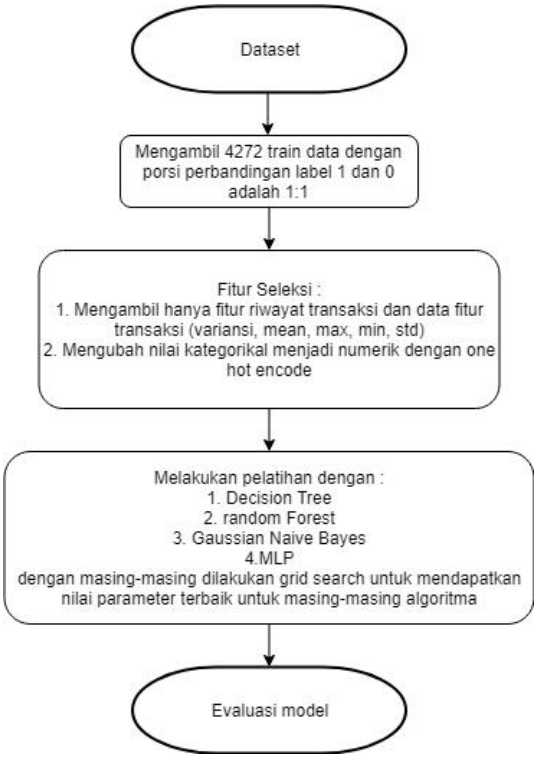
(iv) RNN

Recurrent Neural Network (RNN) adalah jenis dari jaringan syaraf tiruan yang banyak digunakan dalam *speech recognition* dan *Natural Language Processing* (NLP)[5]. RNN dirancang untuk mengenali karakteristik sekuensial data dan menggunakan pola untuk memprediksi kemungkinan berikutnya. Dapat digunakan untuk kasus dimana konteks sangat penting untuk memprediksi hasil akhir. Penggunaan RNN terhubung pada *language model* dimana model tersebut mengetahui huruf berikutnya dalam kata atau kata berikutnya dalam kalimat berdasarkan data sebelumnya.

III. PENGUJIAN DAN PEMBAHASAN

Dalam permasalahan ini, data set \akan digunakan untuk mengklasifikasikan target dalam hal ini adalah gender yang berupa dataset status tweet pada user twitter untuk dapat menentukan apakah tweet tersebut dilakukan oleh user perempuan atau laki-laki. Hal tersebut seperti yang telah dijelaskan di atas disebut sebagai *author profiling*.

Gambar III.1 Flowchart



Sebelum menggunakan data yang diberikan oleh PAN 2016, perlu dilakukan pengaksesan API twitter untuk dapat mengambil data twitter. Dataset dalam bentuk .xml setelah berhasil diambil data tweet dari twitter kemudian ditampung dalam satu file .csv untuk memudahkan pemrosesan data. Hal ini dilakukan karena seterusnya digunakan representasi data dalam bentuk dataframe untuk memudahkan proses klasifikasi.

Kemudian informasi kolom index serta umur di drop sehingga didapatkan dataframe dengan dua kolom yaitu kolom status dan gender sebagai target.

A. Data Cleaning dan Preprocessing

Sebelum melakukan training data, perlu dilakukan *data cleaning* dan *preprocessing* karena tidak semua data dalam keadaan yang baik untuk langsung diproses. Terdapat beberapa data yang perlu dilakukan *preprocessing* untuk meningkatkan kualitas data, sehingga ketika terjadi proses data diolah, akan didapatkan model yang baik. Sehingga dengan diperoleh data dengan kualitas baik akan menghasilkan output yang baik pula. *Data cleaning* yang dilakukan untuk data dalam penelitian ini adalah sebagai berikut:

Gambar III.2 Data Cleaning

```
from bs4 import BeautifulSoup
import re
from nltk.tokenize import WordPunctTokenizer
from nltk.corpus import stopwords

tok = WordPunctTokenizer()
pat1 = r'@[A-Za-z0-9]+'
pat2 = r'https?://[A-Za-z0-9./]+'
combined_pat = r'|'.join((pat1, pat2))
stop_words = set(stopwords.words('english'))
```

1. HTML Encoding

   Digunakan library BeautifulSoup untuk melakukan HTML encoding.
2. Menghilangkan @mention
3. Menghilangkan URL Links

   Digunakan regex untuk melaukan langkah 2 dan 3.
4. Decoding text ke UTF-8
5. Meghilangkan stopwords

Setelah proses *data cleaning* dan *preprocessing* selesai, dilakukan pengecekan nilai NaN yang kemungkinan akan terjadi. Jika ditemukan dilakukan *dropna* untuk menghapus data yang berisi nilai NaN. Pada proses yang dilakukan, dari 5003 data didapatkan 27 data NaN sehingga data yang akan digunakan menjadi 4976 data.

B. Feature Selection

Untuk memperoleh model data yang lebih baik, diperlukan penyeleksian fitur data sebelum melakukan training untuk mencegah terjadinya overfitting. Metode yang dilakukan dalam melakukan *feature selection* adalah menggunakan K-Best Feature Selection dengan Chi-square. Diperoleh10 terms yang paling signifikan adalah *travel , asia, lp, wire, via, great, booksforall, fabrik, week* dan *us.*

Setelah dilakukan *data cleaning, preprocessing* dan *feature selection,* maka data telah siap untuk masuk ke dalam tahap pelatihan data. Pelatiahan data yang dilakukan di penelitian ini adalah dengan menggunakan : 1)Decision Tree; 2)Random Forest; 3)MLP; 4) SVM

IV. HASIL PENGUJIAN DAN PEMBAHASAN

Proses pelatihan dilakukan dengan menggunakan data pelatihan yang telah dilakukan *preprocessing* yang telah dijelaskan pada bab III. Metode digunakan secara berurutan pada dataset yang sama. Scoring yang diterapkan pada masing-masing model dilakukan dengan menggunakan nilai akurasi. Nilai akurasi digunakan untuk mengetahui proporsi benar yang dihasilkan dari sebuah klasifier. Tabel IV.1 menunjukkan hasil akurasi dari masing-masing algoritma klasifikasi yang digunakan.

Algoritma	Akurasi
Decission Tree	0.5147058823529411
Random Forest	0.5110294117647058
MLP	0.6983965902171707
SVM	0.7201136594276436

V. KESIMPULAN

Berdasarkan metode yang telah dilakukan, didapatkan akurasi terbaik yaitu 0.7201 dengan menggunakan classifier SVM. Diperlukan preprocessing data terlebih dahulu untuk menghasilkan data yang lebih bersih dan mudah untuk dilakuakn pelatihan data.

REFERENCES

[1] Deering B.J. (2002) Chapter 11: KM for competitive advantage: mining diverse sources for marketing intelligence. Knowledge Management Strategy and Technology. Bellaver R.F. & Lusa J.M. Editors. Artech House

[2] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Introduction to Data Mining." (2006).

[3] Thapliyal, M.P. Data Mining : A Tool for Banking Industry. April 2015 from : https://www.ermt.net/docs/papers/Volume_4/4_April2015/V4N4-175.pdf

[4] Decision Tree - Classification. From :http://www.saedsayad.com /decision tree.htm .

[5] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[6] YUDA, NUGROHO SEPTIAN. "Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro.(Studi Kasus: Fakultas Ilmu Komputer Angkatan 2009)." *Skripsi, Fakultas Ilmu Komputer* (2014).

[7] Spencer M, Eickholt J, Cheng J. A deep learning network approach to ab initio protein secondary structure prediction. IEEE Transactions on Comp Biol Bioinformatics 2014.

[8] GANI, Irwan; AMALIA, Siti. ALAT ANALISIS DATA: Aplikasi Statistik untuk Penelituan Bidang Ekonomi dan Sosial. Penerbit Andi, 2015.