Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11757

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of Generative Pre-Trained Transformer 3 Classify-Text in Determining Thesis Supervisor

Yoga Handoko Agustin[1)*], Ridwan Setiawan[2)], Iik Abdul Kholik[3)], Wahyu Sindu Prasetya[4)]
[1,2,3)]Institut Teknologi Garut, [4]STMIK Pontianak, Indonesia
[1)]yoga.handoko@itg.ac.id, [2)]ridwan.setiawan@itg.ac.id, [3)]1706103@itg.ac.id, [4)]wahyusinduprasetya@email.com

Submitted : Aug 31, 2022 | **Accepted** : Oct 2, 2022w | **Published** : Oct 3, 2022

Abstract: One of the requirements for graduating from the undergraduate level for universities in Indonesia is writing a final project or thesis. In order to graduate, of course, it is greatly influenced by the desire and strong spirit of the students and also the guidance of the supervisor. In determining the supervising lecturer, special attention must be paid to the field. Usually the selection of lecturers for thesis supervisors is determined by the study program through a meeting of lecturers in order to determine which lecturers are considered according to the title of the student and in accordance with the research of the supervisor. However, this method is a bit inconvenient and also quite time-consuming considering the number of students is more than a hundred and continues to grow every year. In this study, the thesis supervisor was classified based on the title proposed by the student. The methodology that will be used in this research is the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology whose stages are: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and deployment, as well as using Generative Pre-Technology. trained Transformers 3 (GPT-3)

Keywords: CRIPS-DM, GPT-3, Title, Thesis Supervisor

INTRODUCTION

Currently the development of information technology is very fast, every year, month, week, day, hour, minute, even second, technology is discovered or developed. Of course all of this develops according to what is needed by us as humans who almost all fields require technology ranging from industry, education, government, and even agriculture also need technology. From the statement above, it is stated that education is a field that requires technology. Education is a necessity for the people of Indonesia to change humans from being helpless to life into human beings who are efficient, and are expected to be able to produce quality human resources that can contribute to a dignified Indonesian nation (Japar et al., 2018). The level of education in Indonesia is divided into four stages, namely early childhood, elementary, middle, and high. At the age of 3 to 4 or 5 years, they enter early childhood education or kindergarten, continue to elementary school for 6 years, then continue to secondary education, namely 3 years for junior high school and 3 years for high school, and the last one is college(Sesiomadika & 2019, n.d.).

One of the universities in Indonesia is the Garut Institute of Technology which is located in the Garut area of West Java. Garut Institute of Technology is a higher education institution that always produces good graduates and every year continues to increase(Sintiani et al., 2017). One of the requirements to graduate from the Garut Institute of Technology and all universities is to have completed a thesis course. Thesis is a requirement to be able to get a bachelor's degree or a bachelor's degree in all universities in Indonesia(Teknologi & 2017, n.d.). To help students complete their thesis assignments, a supervisor is needed who is in charge of providing constructive direction in determining the title, methodology, writing format and so on (Pendidikan & 2019, n.d.). Usually, the determination of the thesis supervisor is determined by the study program through a lecturer meeting to determine which lecturer is in accordance with the title proposed by the student. However, this method is a bit inconvenient and also quite time-consuming considering the number of students who are not small(Abdullah et al., n.d.).

The application of data mining is needed in order to be able to find a relationship or pattern and trend from large data sets using statistical and mathematical techniques.(Listiani et al., n.d.). One of the data mining techniques is classification, classification is a job to assess a data object, then put it into a certain class from the

*name of corresponding author

number of classes that are already available. Classification performs modeling based on existing training data, then the model is used to classify it on the new data. Classification can be interpreted as an activity of conducting training (training) on the target function that maps each set of attributes or features into one number of class labels that are already available (Utomo et al., 2020).

Several previous studies related to this research, the first entitled "Detecting Hate Speech with GPT-3." Sophisticated language models such as GPT-3 OpenAI can produce hate speech texts that target marginalized groups. The researcher uses GPT-3 to identify sexist and racist parts of the text using the zero-, one-, and few-shot learning methods. The next research is entitled "Computer-Based Decision Making With Simple Multi Attribute Rating Technique Method in Determining Supervisors". This research was made a system with a decision-making technique using the Simple Multi Attribute Rating Technique (SMART) method with the aim of providing a supervisor recommendation for each proposed title.(Putra & Djasmayena, 2020).

Based on the background of the problem above, this study will discuss the Implementation of Generative Pre-trained Transformer 3 (GPT-3) Classify-Text Technology in Determining Thesis Supervisor Based on the Title Proposed by Students.

METHOD

Methods and problem solving in data mining that are commonly used in the world of business and research are using the CRISP-DM method. The stages in this methodology consist of six stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment(Hasanah et al., 2021). This methodological process consists of 6 stages as shown in Fig. 2.1.

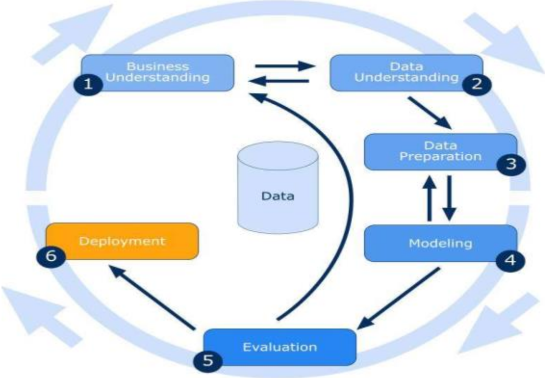


Fig. 2.1. Phases of the CRISP-DM Reference Model

The following is an illustration and a brief description of each phase in CRISP-DM:

1. Business Understanding

   Several things are done at this stage such as understanding the needs and goals from a business point of view then interpreting knowledge into the form of defining problems in data mining and then determining plans and strategies to achieve data mining goals;

2. Data Understanding

   At this stage, it begins with collecting data, then describing the data, then evaluating the quality of the data;

3. *Data* Preparation

   At this stage, the final dataset is created from the raw data that has been prepared earlier. Things that can be done in this stage are data cleaning or data cleaning, then data selection or data selection, attributes, and records, also transformations can be carried out on data which can later be used as suggestions or input in the modeling stage, or can be called Data Transformation;

4. Modeling

   In this stage, machine learning will be involved directly which will later be useful for determining algorithms from data mining, data mining techniques, and tools for data mining. This research uses the GPT-3 Classify-text technology to be able to determine the thesis supervisor based on the proposed title;

*name of corresponding author

5. Evaluation

   At this evaluation stage, it is done by looking at the results or performance levels obtained by the algorithm. Confusion Matrix is one of the benchmarks or parameters that will be used to evaluate the results of this algorithm modeling, by looking at the rules of precision, recall, and accuracy values.

   The confusion matrix can provide an assessment of the performance of the classification model based on the number of objects that are predicted correctly and incorrectly in order to obtain accuracy values and others(Nugraha et al., n.d.). Confusion Matrix is a matrix in the form of a table that serves to record classification performance(Prasetyowati, 2017). The problem solved in the Confusion Matrix is binary classification for two classes. With the Confusion Matrix, the performance of the classification system can be measured properly;

6. Deployment

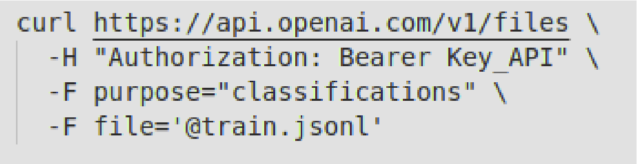
   This stage is done by making reports and journal articles using the resulting model.

RESULT

After passing the CRISP-DM phase, there is a csv file containing 824 datasets of journal titles along with the names of their lecturers obtained from the published journals of each lecturer, then the data is reformatted by translating data from foreign languages into Indonesian, and change the file form from .csv format to JavaScript Object Notation (JSON) Lines or .jsonl format where the text is the title of the research from the lecturer and the label is the name of the lecturer himself. After the data has been collected, classification is then carried out to determine the thesis supervisor based on the title submitted by the student using the GPT-3 Classify-text technology with the following steps:

7. Uploading *labeled examples*

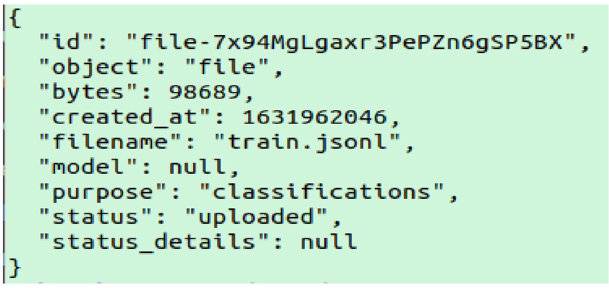
   The data that has been converted into .jsonl format with the label of each lecturer is then uploaded to OpenAI via the terminal in the manner shown in Fig. 3.1



```
curl https://api.openai.com/v1/files \
  -H "Authorization: Bearer Key_API" \
  -F purpose="classifications" \
  -F file='@train.jsonl'
```

Fig 3.1 JSON Lines File Upload Script

   Then the response from the script above is the id of the file which will later be used in modeling classify-text, model, status, purpose, and others. For details, it is listed in Fig 3.2



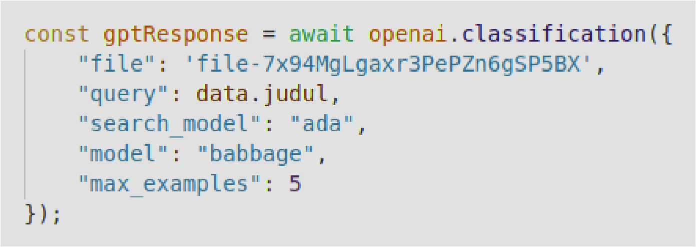
```
{
  "id": "file-7x94MgLgaxr3PePZn6gSP5BX",
  "object": "file",
  "bytes": 98689,
  "created_at": 1631962046,
  "filename": "train.jsonl",
  "model": null,
  "purpose": "classifications",
  "status": "uploaded",
  "status_details": null
}
```

Fig 3.2 Upload Script Response

8. Querying *classifications*

   At this stage, after the data preparation is complete, in order to determine the recommendation for the thesis supervisor, you must access the OpenAI GPT-3 Classiffy-Text Application Programming Interface (API) via the backend. There are several APIs that can access OpenAI which are registered on the official OpenAI website. What I will use is from Nikita Jerschow, for the API it will be explained in Figure 3.3

*name of corresponding author



```
const gptResponse = await openai.classification({
    "file": 'file-7x94MgLgaxr3PePZn6gSP5BX',
    "query": data.judul,
    "search_model": "ada",
    "model": "babbage",
    "max_examples": 5
});
```

Fig 3.3 API OpenAI GPT-3 Classify-Text

In Figure 3.3, it can be seen that for the API from OpenAI there is some data needed to carry out the process of determining the thesis supervisor. The first is "file", this file contains the dataset that has been converted into .jsonl format in the previous stage. The second "query", the query is the data to be trained or classified, namely the title submitted by the student. For "search_model" and "model" both are engines that are used, there are 4 types of engines provided by OpenAI but why do researchers choose "ada" and "babbage" because they are cheaper and the results after trying and comparing are not much different. For "max_example" is the limit of the results that have been obtained, the researcher limits it to 5 maximum results.

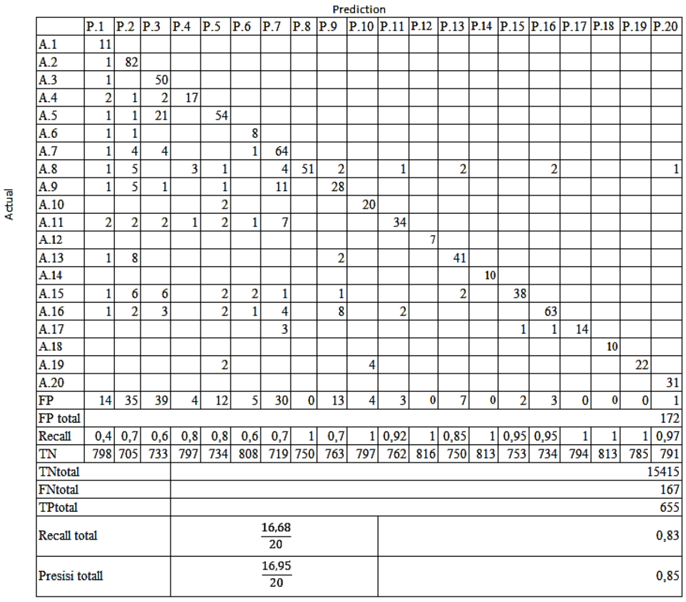
9. *Estimating probabilities of classification labels*

By setting the logprobs parameter and processing the top_logprobs returned in the result, it is possible to estimate the predicted probability of each classification label. There are a few tips on how to format labels for better probability estimation:

a. All labels must be written in capital letters. Otherwise, the input label will be formatted on the backend;
b. Each label must begin with a whitespace when encoded by the tokenizer;
c. It is highly recommended that each label is a single token word. If each label is a single token, the calculated probability will be exact. Otherwise, the probabilities are approximate.

In order to measure the success rate of this research, an external evaluation model is used, namely the Confusion matrix. Confusion Matrix is a tool or technology that can be used to measure classification performance. Confusion Matrix is a matrix in the form of a table that serves to record classification performance. The problem solved in the Confusion Matrix is binary classification for two classes. With the Confusion Matrix, the performance of the classification system can be measured properly(Prasetyowati, 2017). Evaluation using the Confusion matrix can produce precision values, recall, and of course accuracy. Accuracy referred to in classification is the percentage of record accuracy of data that is classified correctly after testing the classification results.

Table 1. Confusion Matrix



Prediction

Actual

	P.1	P.2	P.3	P.4	P.5	P.6	P.7	P.8	P.9	P.10	P.11	P.12	P.13	P.14	P.15	P.16	P.17	P.18	P.19	P.20
A.1	11																			
A.2	1	82																		
A.3	1		50																	
A.4	2	1	2	17																
A.5	1	1	21		54															
A.6	1	1				8														
A.7	1	4	4			1	64													
A.8	1	5		3	1		4	51	2		1		2			2				1
A.9	1	5	1		1		11		28											
A.10					2					20										
A.11	2	2	2	1	2	1	7				34									
A.12												7								
A.13	1	8							2				41							
A.14														10						
A.15	1	6	6		2	2	1		1				2		38					
A.16	1	2	3		2	1	4		8		2					63				
A.17							3								1	1	14			
A.18																		10		
A.19					2					4									22	
A.20																				31
FP	14	35	39	4	12	5	30	0	13	4	3	0	7	0	2	3	0	0	0	1
FP total	172																			
Recall	0,4	0,7	0,6	0,8	0,8	0,6	0,7	1	0,7	1	0,92	1	0,85	1	0,95	0,95	1	1	1	0,97
TN	798	705	733	797	734	808	719	750	763	797	762	816	750	813	753	734	794	813	785	791
TNtotal			15415																	
FNtotal			167																	
TPtotal			655																	
Recall total			$\frac{16,68}{20}$							0,83										
Presisi totall			$\frac{16,95}{20}$							0,85										

*name of corresponding author

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

$$Accuracy = \frac{(655+15415)}{(655+172+167+16415)}$$

$$Accuracy = 0{,}98$$

Accuracy can be said to be perfect if the accuracy reaches 1.00 and is said to be bad if the accuracy is below 0.50 (Agustin et al., n.d.). From table 3.1 it can be seen that the evaluation results using the confusion matrix produce a value of 0.98 or an accuracy of 98%.

DISCUSSION

In the implementation of the GPT-3 Classify-text, the thesis supervisor based on the title submitted by the student can add attributes other than the name and title of the research, by increasing concentration in order to determine better accuracy. As well as being able to try OpenAI engines other than Babbage and exist to produce more varied results

CONCLUSION

The accuracy resulting from the dataset processing of 823 is able to produce an accuracy value of 98% by implementing the GPT-3 Classify-Text algorithm in recommending thesis supervisors. So it is suitable to be implemented in a thesis supervisor decision support system. With GPT-3 Classify-text technology, it can help in the distribution of supervisors that are relevant to the title submitted by students.

REFERENCES

Abdullah, A., dan, M. P.-J. E., & 2018, undefined. (n.d.). Rancang Bangun Sistem Pendukung Keputusan Dalam Pemilihan Dosen Pembimbing Skripsi Dengan Metode AHP di UM Pontianak. *Download.Garuda.Kemdikbud.Go.Id*. Retrieved August 30, 2022, from http://download.garuda.kemdikbud.go.id/article.php?article=934062&val=13360&title=Rancang%20Bangun%20Sistem%20Pendukung%20Keputusan%20Dalam%20Pemilihan%20Dosen%20Pembimbing%20Skripsi%20Dengan%20Metode%20AHP%20di%20UM%20Pontianak

Agustin, Y., Kusrini, K., … E. L.-S. R. and, & 2017, undefined. (n.d.). Klasifikasi Penerimaan Mahasiswa Baru Menggunakan Algortima C4. 5 Dan Adaboost (Studi Kasus: STMIK XYZ). *Csrid.Potensi-Utama.Ac.Id*. Retrieved August 30, 2022, from http://csrid.potensi-utama.ac.id/ojs/index.php/CSRID/article/view/126

Hasanah, M., … S. S.-J. of A., & 2021, undefined. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Jurnal.Polibatam.Ac.Id*, *5*(2), 103. https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/3200

Japar, M., Zulela, M., & Mustoip, S. (2018). *Implementasi Pendidikan Karakter*. https://books.google.com/books?hl=id&lr=&id=OqB_DwAAQBAJ&oi=fnd&pg=PA1&dq=Implementasi+Pendidikan+Karakter+-+Muhammad+Japar,+Zulela+MS,,+Sofyan+Mustoip,&ots=UMNKi8zyFK&sig=2RvO8_6CVVwOUowdHlRvJjlLXJ0

Listiani, L., … Y. A.-… S. I. dan, & 2019, undefined. (n.d.). Implementasi algoritma k-means cluster untuk rekomendasi pekerjaan berdasarkan pengelompokkan data penduduk. *Ejurnal.Dipanegara.Ac.Id*. Retrieved August 30, 2022, from https://ejurnal.dipanegara.ac.id/index.php/sensitif/article/view/439

Nugraha, W., Com, R. S.-Techno., & 2021, undefined. (n.d.). Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4. 5, Random Forest, dan SVM. *Publikasi.Dinus.Ac.Id*. Retrieved August 30, 2022, from http://publikasi.dinus.ac.id/index.php/technoc/article/view/4762

Pendidikan, A. A.-K., & 2019, undefined. (n.d.). Hubungan Antara Capaian Pembelajaran Dasar-Dasar Penelitian Dan Statistik Dengan Mutu Skripsi: Studi Analisis di STKIP Muhammadiyah Bogor. *Jurnalnasional.Ump.Ac.Id*. Retrieved August 30, 2022, from http://jurnalnasional.ump.ac.id/index.php/khazanah/article/download/4290/2495

Prasetyowati, E. (2017). *DATA MINING Pengelompokan Data untuk Informasi dan Evaluasi*. https://books.google.com/books?hl=id&lr=&id=rEH2DwAAQBAJ&oi=fnd&pg=PA4&dq=E.Prasetyowati,%E2%80%9CDATA+MINING+Pengelompokan+Data+untuk+Informasi+dan+Evaluasi,%E2%80%9D+2017,+Accessed:+Jul.+16,+2022.+%5BOnline%5D.+Available:+https://www.google.com/books%3Fhl%3

Did%26lr%3D%26id%3DrEH2DwAAQBAJ%26oi%3Dfnd%26pg%3DPA4%26dq%3DPrasetyowati,%2BE.%2B(2017).%2BDATA%2BMINING%2BPengelompokan&ots=kw6BDG37r8&sig=A05lgfJ_d8PrXhCfMy-STCwfF00

Putra, R. E., & Djasmayena, S. (2020). Metode Simple Multi Attribute Rating Technique Dalam Keputusan Pemilihan Dosen Berprestasi yang Tepat. *Jurnal Informasi & Teknologi*, *2*(1). https://doi.org/10.37034/JIDT.V2I1.29

Sesiomadika, S. N.-P., & 2019, undefined. (n.d.). APLIKASI SUPREMUM DAN INFIMUM DALAM KEBIJAKAN PENDIDIKAN INDONESIA. *Journal.Unsika.Ac.Id*. Retrieved August 30, 2022, from https://journal.unsika.ac.id/index.php/sesiomadika/article/view/2116

Sintiani, I., Algoritma, L. F.-J., & 2017, undefined. (2017). Pengembangan Aplikasi Tracer Study STT-Garut. *Jurnal.Sttgarut.Ac.Id*. http://jurnal.sttgarut.ac.id/index.php/algoritma/article/view/461

Teknologi, M. Y.-J. T. T. I. I., & 2017, undefined. (n.d.). Penerapan jaringan syaraf tiruan dengan algoritma perceptron pada pola penentuan nilai status kelulusan sidang skripsi. *Teknoif.Itp.Ac.Id*. Retrieved August 30, 2022, from https://www.teknoif.itp.ac.id/index.php/teknoif/article/view/178

Utomo, D., Informatika, M. M.-J. M., & 2020, undefined. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Ejurnal.Stmik-Budidarma.Ac.Id*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080