6th Semester, Academic Year 2026-27

GEN-AI

HANDSON_2

Date: 28/1/26

Name: N S LIKHITH CHANDRA	SRN: PES2UG23CS366	Section:F

Problem Statement :

Legal and policy documents are lengthy and complex for users to understand. This project aims to summarize such documents and extract important entities to improve readability and awareness.

Abstract :

The system integrates text summarization and named entity recognition using pretrained transformer models from Hugging Face to simplify and analyze policy documents efficiently.

Short Documentation

- In this project, I learned how pretrained NLP models from Hugging Face can be used directly through pipelines without doing any model training.
- How text summarization works and how a large document can be converted into a short and meaningful summary using transformer-based models.
- built a system that takes a long privacy policy document as input and produces a simplified summary so that users can quickly understand the main points.
- Along with summarization, implemented Named Entity Recognition (NER) to identify important information such as organization names, locations, and dates from the document.
- multiple NLP tasks can be combined together in a single workflow to solve a real-world problem.
- Overall, how Hugging Face pipelines can be used to build practical NLP applications with minimal code.

INPUT → OUTPUT :

Input

- Long legal / policy text (single .txt file)

Processing

1. Summarization pipeline
2. NER pipeline

Output

- Simplified policy summary
- List of extracted entities:
    - Organizations
    - Dates
    - Legal references

Entities extracted (ss) :

```
entities = ner(text)

for ent in entities:
    print(ent['word'], "->", ent['entity_group'])
```

```
this privacy policy explains how -> LABEL_0
abc -> LABEL_1
technologies pv -> LABEL_0
##t ltd -> LABEL_1
collects, uses, stores, and protects user information when individuals use our website and services. we collect personal information such as name, email address, -> LABEL_0
contact number -> LABEL_1
, and usage data when users register, submit forms, or interact with our platform. this information is collected to provide better services, personalize user -> LABEL_0
experience -> LABEL_1
, and improve system performance. -> LABEL_0
abc -> LABEL_1
technologies pv -> LABEL_0
##t ltd -> LABEL_1
may also collect non - personal information such as -> LABEL_0
browser -> LABEL_1
type, ip address, operating system, and access -> LABEL_0
times -> LABEL_1
##tamps for analytics and security purposes. cookies may be used to enhance -> LABEL_0
website -> LABEL_1
functionality and track user preferences. user data may be shared with trusted third - party service providers for purposes such as payment processing, customer support, and infrastructure maintenance. these thi
company -> LABEL_1
and its users. all collected data is stored securely using encryption and access controls. we retain personal data only for as long as necessary to fulfill the purposes outlined in this policy unless a -> LABEL_
longer -> LABEL_1
retention period is required by law. users have the -> LABEL_0
right -> LABEL_1
to access, update, or request deletion of their personal data. requests -> LABEL_0
can -> LABEL_1
be -> LABEL_0
submitted -> LABEL_1
by contacting our support team via email. this privacy policy is governed by -> LABEL_0
the -> LABEL_1
laws -> LABEL_0
of -> LABEL_1
```

```
by contacting our support team via email. this privacy policy is governed by -> LABEL_0
the -> LABEL_1
laws -> LABEL_0
of -> LABEL_1
india. -> LABEL_0
any disputes arising -> LABEL_1
under this policy shall be -> LABEL_0
subject -> LABEL_1
to the -> LABEL_0
jurisdiction of the courts of -> LABEL_1
bengal -> LABEL_0
##uru -> LABEL_1
. -> LABEL_0
abc -> LABEL_1
technologies pv -> LABEL_0
##t ltd -> LABEL_1
reserves -> LABEL_0
the right -> LABEL_1
to update or modify this policy at any time. changes will be effective from the date of publication on the -> LABEL_0
website -> LABEL_1
. -> LABEL_0
users are advised -> LABEL_1
to review this -> LABEL_0
policy periodically -> LABEL_1
to -> LABEL_0
stay informed -> LABEL_1
. -> LABEL_0
```

Summary :

```
summary = summarizer(text, max_length=120, min_length=50, do_sample=False)
print("SUMMARY:\n", summary[0]['summary_text'])

SUMMARY:
 ABC Technologies Pvt Ltd collects, uses, stores, and protects user information when individuals use our website and services . We collect personal information such as name, email address, contact number, usage
```

SUMMARY:

ABC Technologies Pvt Ltd collects, uses, stores, and protects user information when individuals use our website and services . We collect personal information such as name, email address, contact number, usage data when users register, submit forms, or interact with our platform . This Privacy Policy is governed by the laws of India .

Observations :

- The summarization model was able to reduce a long privacy policy into a short and meaningful summary while preserving the main ideas.
- The quality of the summary depends on the length and clarity of the input text; very long or complex sentences may lead to minor information loss.
- The DistilBART model generated summaries faster than larger models, making it suitable for CPU-based execution in Google Colab.
- The Named Entity Recognition (NER) model successfully identified organization names, locations, and dates from the policy document.
- Some entities were occasionally misclassified or missed due to the generic nature of the pretrained NER model.
- The system works well without any dataset or model training, demonstrating the effectiveness of pretrained transformer models.

- Combining summarization and NER made the output more informative compared to using only a single NLP task.