Classification to Support Trustworthy Digital Records

AI for Classification

[Gita Sastria, **Umi Asma' Mokhtar**, Sabrina Tiun, Masnizah Mohd]

[3rd ITRUST AI Symposium]
[7th July 2023]

Introduction

- Today's information environments have become a
  - 'wild frontier', or… most trending?
  - decentralised and fractured, and
  - subject to pressures that include increasing data volumes, reliance on commercial and proprietary systems, and evolving forms of records and formats.

InterPARES TrustAI

- Records management is defined as a (ISO 15489-1, 2001).
  - supervision and administration of digital or paper records,
  - regardless of format,
  - responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposal
  - including processes for capturing, and **maintaining evidence** of and information about business activities and transactions in the form of records

InterPARES TrustAI

- Some of the major records management challenges are
  - information overload,
  - data accuracy & integrity,
  - maintaining compliance with regulations,
  - managing records across multiple locations,
  - poor records retrieval, and
  - missing retention schedule

InterPARES TrustAI

- Organizations face unprecedented uncertainty and struggling with the unsustainable task of keeping control and managing overwhelming number of digital records. Hence, drowning in records and information
  - **71%** of organisations have **no idea of the content** in their **stored** data, and **79%** of organisations say too **much time and effort is spent manually** searching and disposing of information .
  - Yet, in 2019, only **44%** of records professionals agreed that their organizations **use automated** tools to locate and preserve relevant information, **49% manually delete** emails and **58% manually delete** records **from mobile devices** (Cohasset-ARMA, 2019).

InterPARES TrustAI

Revisit: Can Machines Classify Better than Humans?

- Exponential increment number of digital records make **manual classification** task become **harder** and more **irrelevant** for human to handle and can make delays in making decisions accessible.
- The archival discipline consists in building knowledge about archival documents and acting upon them in methodical ways to protect the properties that they have.

InterPARES TrustAI

- **Foundational constructs** from the archival theory, including
  - "record" (document, information, data) and
  - "trustworthiness" (reliability, authenticity—identity and integrity, and accuracy), and
  - concepts such as the characteristics of archival documents or records (impartiality, authenticity, naturalness, interrelatedness, and uniqueness; and the network of origin, and determined relationships between and among records (archival bond).

InterPARES TrustAI

How Machines Classify Better than Humans?

- Most of the previous classification works on records content only.
- The complexity of records classification lies on context & structure.
  - **Content:** text, data, symbol, image, sound, graphic, and any information forming a record
  - **Context:** 3 aspects are contextual information (e.g., digital signature), relationship from record to another record, and activity that create the record. - E.g., via doc/info attached, network, reference/code/number or metadata
  - **Structure:** how records are formed including: format, symbol e.g., letter, memo, official email in public office; address, date, paragraph, and signature.

InterPARES TrustAI

Method (1) Text Mining

- Text mining is the process of transforming **unstructured text into a structured format to identify meaningful patterns and new insights** by applying advanced analytical techniques such as Naïve Bayes, Support Vector Machines (SVM), and others AI techniques.
- *A document or record consists of many unstructured data that can be explore and discover hidden relationships for new novel of knowledge and information.*
- Since roughly 80% of data in the world resides in an unstructured format, text mining is an extremely valuable practice within organizations.
- There are several text mining techniques,
  - Information Retrieval
  - Natural Language Processing
  - Information Extraction
  - Data Mining

InterPARES TrustAI

- Text mining:
  - was successfully implemented on many texts problems analysis and bringing a new insight on text classification area with outperform performance.
  - Most of the problems solved using text mining techniques use data sets in the form of short text.
- *The use of short text data sets provides an advantage because the amount of text used are very limited and make algorithm can work very fast and stable.*

- In this research we used text mining techniques for classification and arrange records based on their function and observe text mining algorithm performance when handle long text data sets like documents or records.

InterPARES TrustAI

- Several text mining techniques will be used for classification and arrangement of records based on function for content, context and structure of records

Content classification	Context classification:	Structure classification
Text Categorization and **Text Summarization** techniques will be used to extract knowledge from the text.	**Named-entity Recognition** technique will be adopted to extract contextual information from records.	**Knowledge Graph** technique will be proposed to build relationship among records

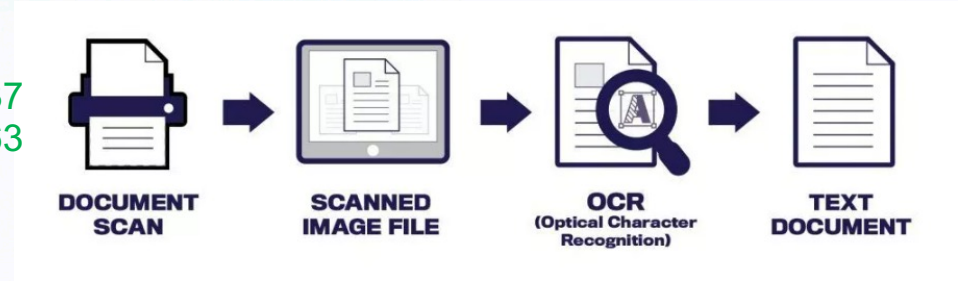
InterPARES TrustAI

PRE-LIM EXPERIMENT

(1) DATA COLLECTION

- We collect 500 real data, from the University. The data source is correspondence at the faculty in the 2022.
- The correspondence data has been arranged based on:
  - the date of issued
  - type of incoming and outgoing correspondence since it was collected.

Incoming letter: 237
Outgoing letter: 463



CR technique: Tesseract library

Text doc verification – using Dictionary (2008)

Use Boyer Moore string pattern matching algo with Jaro Winkler to verify unmatching words collection with the Dictionary:

Text documents with minimum errors have been generated through the previous phase, the texts from the unstructured documents will be extracted and established to produce metadata.

InterPARES TrustAI

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN,
RISET DAN TEKNOLOGI
UNIVERSITAS RIAU
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
JURUSAN MATEMATIKA
Kampus Bina Widya Km. 12,5 Simpang Baru Pekanbaru 28293 Telp. (0761) 63273

Nomor : 12956/UN19.5.1.1.3/TU.00.01/2023 — 13 Juni 2023
Lampiran : Satu berkas
Hal : Izin Penelitian dan Pengambilan Data

Yth. Dekan FMIPA
Di Universitas Riau

Sehubungan dengan pelaksanaan Skripsi sebagai Tugas Akhir di Jurusan Matematika FMIPA Universitas Riau, maka dengan ini kami mohon kepada Bapak untuk dapat membuatkan surat pengantar izin penelitian dan pengambilan data untuk mahasiswa sebagai berikut :

Nama	:	Adila Mutiah Arja
NIM	:	1803113133
Program Studi	:	S1 Statistika
Judul Skripsi	:	Analisis Metode Analisis Data *Envelopment* Pada Pengukuran Efisiensi Kinerja Program Studi Di FMIPA Universitas Riau
Tujuan Surat	:	Dekan FMIPA Universitas Riau Cq. Sub. Koordinator Akademik

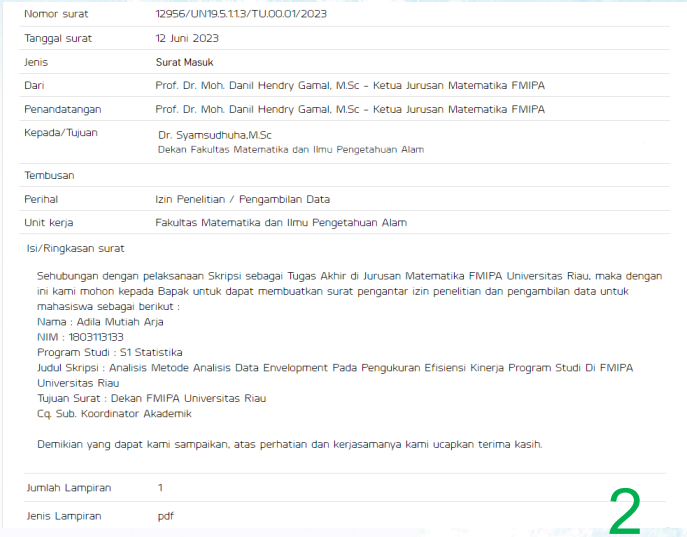
Demikian yang dapat kami sampaikan, atas perhatian dan kerjasamanya kami ucapkan terima kasih.

Prof. Dr. M. D. H. Gamal, M.Sc
NIP. 196506041991031002

1

Example of Incoming Correspondence Type

Related data will be collected from the correspondence such as the registration number, date, type, owner, purpose, copy, subject matter, work unit content or summary of the correspondence, number of attachments, and kind of attachments to be utilized as metadata.



Nomor surat	12956/UN19.5.1.1.3/TU.00.01/2023
Tanggal surat	12 Juni 2023
Jenis	Surat Masuk
Dari	Prof. Dr. Moh. Danil Hendry Gamal, M.Sc - Ketua Jurusan Matematika FMIPA
Penandatangan	Prof. Dr. Moh. Danil Hendry Gamal, M.Sc - Ketua Jurusan Matematika FMIPA
Kepada/Tujuan	Dr. Syamsudhuha,M.Sc Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Tembusan	
Perihal	Izin Penelitian / Pengambilan Data
Unit kerja	Fakultas Matematika dan Ilmu Pengetahuan Alam

Isi/Ringkasan surat

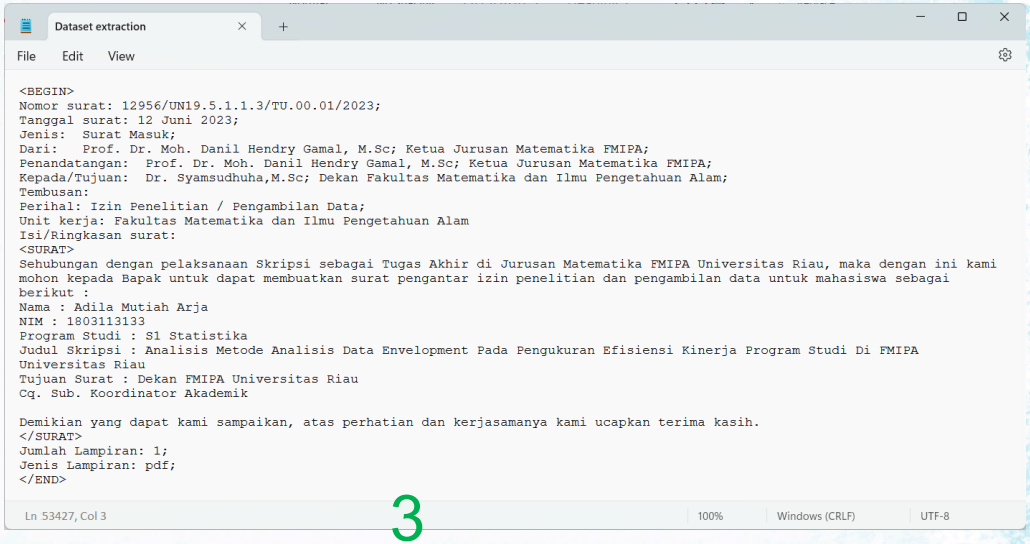
Sehubungan dengan pelaksanaan Skripsi sebagai Tugas Akhir di Jurusan Matematika FMIPA Universitas Riau, maka dengan ini kami mohon kepada Bapak untuk dapat membuatkan surat pengantar izin penelitian dan pengambilan data untuk mahasiswa sebagai berikut :
Nama : Adila Mutiah Arja
NIM : 1803113133
Program Studi : S1 Statistika
Judul Skripsi : Analisis Metode Analisis Data Envelopment Pada Pengukuran Efisiensi Kinerja Program Studi Di FMIPA Universitas Riau
Tujuan Surat : Dekan FMIPA Universitas Riau
Cq. Sub. Koordinator Akademik

Demikian yang dapat kami sampaikan, atas perhatian dan kerjasamanya kami ucapkan terima kasih.

Jumlah Lampiran	1
Jenis Lampiran	pdf

2

Example of a visualization of information extraction



Dataset extraction

File Edit View

```
<BEGIN>
Nomor surat: 12956/UN19.5.1.1.3/TU.00.01/2023;
Tanggal surat: 12 Juni 2023;
Jenis:  Surat Masuk;
Dari:   Prof. Dr. Moh. Danil Hendry Gamal, M.Sc; Ketua Jurusan Matematika FMIPA;
Penandatangan:  Prof. Dr. Moh. Danil Hendry Gamal, M.Sc; Ketua Jurusan Matematika FMIPA;
Kepada/Tujuan:  Dr. Syamsudhuha,M.Sc; Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam;
Tembusan:
Perihal: Izin Penelitian / Pengambilan Data;
Unit kerja: Fakultas Matematika dan Ilmu Pengetahuan Alam
Isi/Ringkasan surat:
<SURAT>
Sehubungan dengan pelaksanaan Skripsi sebagai Tugas Akhir di Jurusan Matematika FMIPA Universitas Riau, maka dengan ini kami
mohon kepada Bapak untuk dapat membuatkan surat pengantar izin penelitian dan pengambilan data untuk mahasiswa sebagai
berikut :
Nama : Adila Mutiah Arja
NIM : 1803113133
Program Studi : S1 Statistika
Judul Skripsi : Analisis Metode Analisis Data Envelopment Pada Pengukuran Efisiensi Kinerja Program Studi Di FMIPA
Universitas Riau
Tujuan Surat : Dekan FMIPA Universitas Riau
Cq. Sub. Koordinator Akademik

Demikian yang dapat kami sampaikan, atas perhatian dan kerjasamanya kami ucapkan terima kasih.
</SURAT>
Jumlah Lampiran: 1;
Jenis Lampiran: pdf;
</END>
```

Ln 53427, Col 3 — 100% — Windows (CRLF) — UTF-8

3

Data will be stored in text file

Metadata information will be used for further phases, beginning with the establishment of a correspondence classification model, and advancing through the preparation of records.

InterPARES TrustAI

(2) LABELLING, CLASSIFYING AND EXTRACTING CORRESPONDENCE TEXT

- Letters correspondence will be labelled and classified **manually** into incoming and outgoing letter types
  - Incoming letters are all types of letters received from outside the organization,
  - Outgoing letters are any sort of letter made and issued to a group inside the organization or another agency or person outside the organization

- All text, especially the content of the letters,
  - will be extracted using the OCR technique and
  - transform and load them into data warehouse and used as a primary dataset.

InterPARES TrustAI

(3) PRE-PROCESSING TEXT

- Correspondence has enormous number of texts, it is important to reduce the text into pieces that are important for the classification or others text mining techniques.
  - Challenge: The data collected might contain portions that cannot be used for the classification and could make the classification more challenging.
  - Text pre-processing is concerned with cleaning the data and removing the noise and unwanted pieces in the text
- Next, will process into groups of single words/token

Text processing

- Transformation
- Tokenization
- Removal of stop words
- Lemmatization & Stemming
- POS Tagging

Classification

- normally achieved in two main ways, i.e., machine learning techniques and lexicon based (score-based) approaches.

Testing & Evaluation

- Text classification performed will be evaluated using formulas in accuracy, precision, recall and F-measure.

- Transformation
  - step removing unnecessary words from the text.
  - minimize noise and improve classification accuracy

```
#case folding
text= text.lower()
```

example output of transformation phase

Before	After
Sehubungan dengan pelaksanaan Skripsi sebagai Tugas Akhir di Jurusan Matematika FMIPA Universitas Riau, maka dengan ini kami mohon kepada Bapak untuk dapat membuatkan surat pengantar izin penelitian dan pengambilan data untuk mahasiswa sebagai berikut : Nama : Adila Mutiah Arja NIM : 1803113133 Program Studi : S1 Statistika Judul Skripsi : Analisis Metode Analisis Data Envelopment Pada Pengukuran Efisiensi Kinerja Program Studi Di FMIPA Universitas Riau Tujuan Surat : Dekan FMIPA Universitas Riau Cq. Sub. Koordinator Akademik Demikian yang dapat kami sampaikan, atas perhatian dan kerjasamanya kami ucapkan terima kasih.	sehubungan dengan pelaksanaan skripsi sebagai tugas akhir di jurusan matematika fmipa universitas riau, maka dengan ini kami mohon kepada bapak untuk dapat membuatkan surat pengantar izin penelitian dan pengambilan data untuk mahasiswa sebagai berikut nama adila mutiah arja nim 1803113133 program studi s1 statistika judul skripsi analisis metode analisis data envelopment pada pengukuran efisiensi kinerja program studi di fmipa universitas riau tujuan surat dekan fmipa universitas riau cq sub koordinator akademik demikian yang dapat kami sampaikan, atas perhatian dan kerjasamanya kami ucapkan terima kasih

Example of Transformation Phase

InterPARES TrustAI

- Tokenization
  - splitting a sentence into words or phrases called tokens.
  - can be identified by punctuation marks or white spaces

```
#tokenizing
text = text.replace('\\t'," ").replace('\\n'," ").replace('\\u'," 
").replace('\\'," ")
text = text.encode('ascii', 'replace').decode('ascii')
text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," 
",text).split())
text = text.replace("http://", " ").replace("https://", " ")
text = text.translate(str.maketrans("","",string.punctuation))
text = re.sub(r"\d+", "", text)
text = text.strip()
text = re.sub('\s+',' ', text)
text = re.sub(r"\b[a-zA-Z]\b", "", text)
```

example output of tokenization phase

Before	After
sehubungan dengan pelaksanaan skripsi sebagai tugas akhir di jurusan matematika fmipa universitas riau, maka dengan ini kami mohon kepada bapak untuk dapat membuatkan surat pengantar izin penelitian dan pengambilan data untuk mahasiswa sebagai berikut nama adila mutiah arja nim 1803113133 program studi s1 statistika judul skripsi analisis metode analisis data envelopment pada pengukuran efisiensi kinerja program studi di fmipa universitas riau tujuan surat dekan fmipa universitas riau cq sub koordinator akademik demikian yang dapat kami sampaikan, atas perhatian dan kerjasamanya kami ucapkan terima kasih	'sehubungan', 'dengan', 'pelaksanaan', 'skripsi', 'sebagai', 'tugas', 'akhir', 'di', 'jurusan', 'matematika', 'fmipa', 'universitas', 'riau,', 'maka', 'dengan', 'ini', 'kami', 'mohon', 'kepada', 'bapak', 'untuk', 'dapat', 'membuatkan', 'surat', 'pengantar', 'izin', 'penelitian', 'dan', 'pengambilan', 'data', 'untuk', 'mahasiswa', 'sebagai', 'berikut', 'nama', 'adila', 'mutiah', 'arja', 'nim', '1803113133', 'program', 'studi', 's1', 'statistika', 'judul', 'skripsi', 'analisis', 'metode', 'analisis', 'data', 'envelopment', 'pada', 'pengukuran', 'efisiensi', 'kinerja', 'program', 'studi', 'di', 'fmipa', 'universitas', 'riau', 'tujuan', 'surat', 'dekan', 'fmipa', 'universitas', 'riau', 'cq', 'sub', 'koordinator', 'akademik', 'demikian', 'yang', 'dapat', 'kami', 'sampaikan,', 'atas', 'perhatian', 'dan', 'kerjasamanya', 'kami', 'ucapkan', 'terima', 'kasih'

Example of Tokenization Phase

- # Removal of Stop Words
  - includes articles and prepositions like ‘the’, ‘and’ ‘or’ ‘still’ ‘which’, etc .
  - usually filtered and removed during pre-processing

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory, StopWordRemover, ArrayDictionary

factory = StopWordRemoverFactory().get_stop_words()
txt_stopword = pd.read_csv('filestopword.txt', names=["stopwords"],
header= None)
txt_stopword= txt_stopword["stopwords"][0].split(' ')
stopword = factory + txt_stopword
dictionary = ArrayDictionary(stopword)
stop_remover=StopWordRemover(dictionary)
text = stop_remover.remove(text)
```

example output

Before	After
'sehubungan', 'dengan', 'pelaksanaan', 'skripsi', 'sebagai', 'tugas', 'akhir', 'di', 'jurusan', 'matematika', 'fmipa', 'universitas', 'riau,', 'maka', 'dengan', 'ini', 'kami', 'mohon', 'kepada', 'bapak', 'untuk', 'dapat', 'membuatkan', 'surat', 'pengantar', 'izin', 'penelitian', 'dan', 'pengambilan', 'data', 'untuk', 'mahasiswa', 'sebagai', 'berikut', 'nama', 'adila', 'mutiah', 'arja', 'nim', '1803113133', 'program', 'studi', 's1', 'statistika', 'judul', 'skripsi', 'analisis', 'metode', 'analisis', 'data', 'envelopment', 'pada', 'pengukuran', 'efisiensi', 'kinerja', 'program', 'studi', 'di', 'fmipa', 'universitas', 'riau', 'tujuan', 'surat', 'dekan', 'fmipa', 'universitas', 'riau', 'cq', 'sub', 'koordinator', 'akademik', 'demikian', 'yang', 'dapat', 'kami', 'sampaikan,', 'atas', 'perhatian', 'dan', 'kerjasamanya', 'kami', 'ucapkan', 'terima', 'kasih'	'sehubungan', 'pelaksanaan', 'skripsi', 'tugas', 'jurusan', 'matematika', 'fmipa', 'universitas', 'riau,', 'mohon', 'membuatkan', 'surat', 'pengantar', 'izin', 'penelitian', 'pengambilan', 'data', 'mahasiswa', 'nama', 'adila', 'mutiah', 'arja', 'nim', '1803113133', 'program', 'studi', 's1', 'statistika', 'judul', 'skripsi', 'analisis', 'metode', 'analisis', 'data', 'envelopment', 'pengukuran', 'efisiensi', 'kinerja', 'program', 'studi', 'fmipa', 'universitas', 'riau', 'tujuan', 'surat', 'dekan', 'fmipa', 'universitas', 'riau', 'cq', 'sub', 'koordinator', 'akademik', 'sampaikan,', 'perhatian', 'kerjasamanya', 'ucapkan', 'terima', 'kasih'

Example

- # Lemmatization and Stemming
  - process of reducing inflectional and other derivations of a word to its dictionary form (root word) called lemma

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

#stemming
factory = StemmerFactory()
stemmer = factory.create_stemmer()
text = stemmer.stem(text)
```

example output

Before	After
'sehubungan', 'pelaksanaan', 'skripsi', 'tugas', 'jurusan', 'matematika', 'fmipa', 'universitas', 'riau,', 'mohon', 'membuatkan', 'surat', 'pengantar', 'izin', 'penelitian', 'pengambilan', 'data', 'mahasiswa', 'nama', 'adila', 'mutiah', 'arja', 'nim', '1803113133', 'program', 'studi', 's1', 'statistika', 'judul', 'skripsi', 'analisis', 'metode', 'analisis', 'data', 'envelopment', 'pengukuran', 'efisiensi', 'kinerja', 'program', 'studi', 'fmipa', 'universitas', 'riau', 'tujuan', 'surat', 'dekan', 'fmipa', 'universitas', 'riau', 'cq', 'sub', 'koordinator', 'akademik', 'sampaikan,', 'perhatian', 'kerjasamanya', 'ucapkan', 'terima', 'kasih'	'hubung', 'laksana', 'skripsi', 'tugas', 'jurus', 'matematika', 'fmipa', 'universitas', 'riau', 'mohon', 'buat', 'surat', 'antar', 'izin', 'teliti', 'ambil', 'data', 'mahasiswa', 'nama', 'adila', 'mutiah', 'arja', 'nim', '1803113133', 'program', 'studi', 's1', 'statistika', 'judul', 'skripsi', 'analisis', 'metode', 'analisis', 'data', 'envelopment', 'ukur', 'efisiensi', 'kerja', 'program', 'studi', 'fmipa', 'universitas', 'riau', 'tuju', 'surat', 'dekan', 'fmipa', 'universitas', 'riau', 'cq', 'sub', 'koordinator', 'akademik', 'sampai', 'perhatian', 'kerjasama', 'ucap', 'terima', 'kasih'

Example

- Part of Speech POS Tagging
  - words can be categorized according to their grammatical properties and syntactic functions called part of speech, such as noun, pronoun, verb, adverb, adjective, conjunction, preposition, and interjection.

Tokens	POS Tagging
hubung	verb
laksana	noun
skripsi	noun
bagai	noun
tugas	noun
akhir	noun

example output

InterPARES TrustAI

(4) CLASSIFICATION AND EVALUATION

- previously cleaned dataset will be separated into 80% training data and 20% testing data. .
- support vector machine approach will be used to train the training data and build model.
- Data testing is used to validate the model

```
#pembagian data training dan data test
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(dt['content'],
dt['label'], test_size=0.2, stratify=dt['label'], random_state = 30)

#membagi data  menjadi data training dan data testing, test_size= 0.2
yang artinya bayaknya data testing adalah 20%
```

1. Script for Separate Train and Testing Dataset

```
#tf idf
from sklearn.feature_extraction.text import TfidfVectorizer
import pickle
vectorizer = TfidfVectorizer()

X_train= vectorizer.fit_transform(X_train)
pickle.dump(vectorizer, open('tfidf6.pkl', 'wb'))
X_test = vectorizer.transform(X_test)
```

2. Tf-Idf Word Weighting

Modelling uses the support vector machine method with kernel parameter as linear

```
#suport vector machine
from sklearn import svm
from sklearn.model_selection import cross_val_score
clf = svm.SVC(kernel = 'linear').fit(X_train,y_train)

#menyimpan model svm
pickle.dump(clf, open('svm6.pkl', 'wb'))

#prediksi data test
prediksi = clf.predict(X_test)
```

3. Build SVM Model

InterPARES TrustAI

The evaluation process uses a confusion matrix to state the amount of test data that is correctly classified and the amount of test data that is misclassified.

```
#evaluasi
from sklearn.metrics import classification_report
print(classification_report(y_test, prediksi))
```

script for display evaluation.

	precision	recall	f1-score	support
-1	0.81	0.85	0.83	50
1	0.90	0.92	0.91	50
accuracy			0.86	100
macro avg	0.86	0.86	0.86	100
weighted avg	0.86	0.86	0.86	100

Confusion Matrix

According to the matrix, the model's prediction accuracy is roughly 86%.

InterPARES TrustAI

(5) ENTITY RECOGNIZE AND ENTITY LINKING

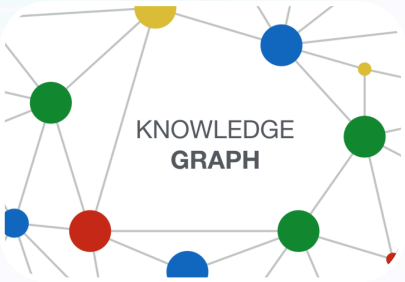
- Entity Recognition is the **process of extracting and identifying essential information from text**.
  - The information that is extracted and categorized is called an entity. It can be any word or a series of words that consistently refer to the same thing.
- Entity linking is the **process of connecting entity mentions in text to their knowledge base counterparts**.
- steps involved to build ER&EL:

InterPARES TrustAI

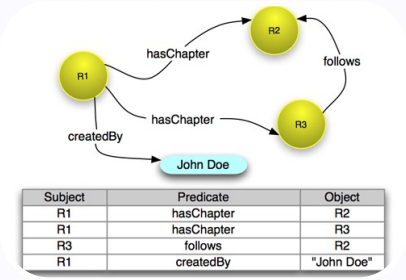
(6) ONTOLOGY

- is an explicit representation of a shared conceptualization, where conceptualization refers to knowledge of an abstract model of a particular domain.
- Ontologies provide necessary semantics in a heterogeneous environment and contribute to information exchange and knowledge discovery.
- Several research activities will be carried out in this phase, including:

Domain definition



Concept identification



Model development

Refinement

InterPARES TrustAI

(7) ORGANIZED AND CLASSIFIED RECORDS

- The entire original data source in letter form will be reorganized based on the knowledge we have extracted from previous steps and built into a collection of records.
- Records will be classified based on function and a classification model generated using knowledge extraction from previous activities.
- We will identify completeness, creation procedure, integrity, identity, precise, correctness, truthfulness, and pertinent.
- With the aid of the extracted knowledge, we create a function-based record classification model based on the **record continuum model** to achieve records trustworthiness.

(8) EVALUATE

Evaluation of the classification results will be performed by domain experts (agency records managers), who will confirm that the correct classify classes have been applied

InterPARES TrustAI

SUBJECT MATTER EXPERTS

InterPARES TrustAI

THANK YOU

InterPARES TrustAI