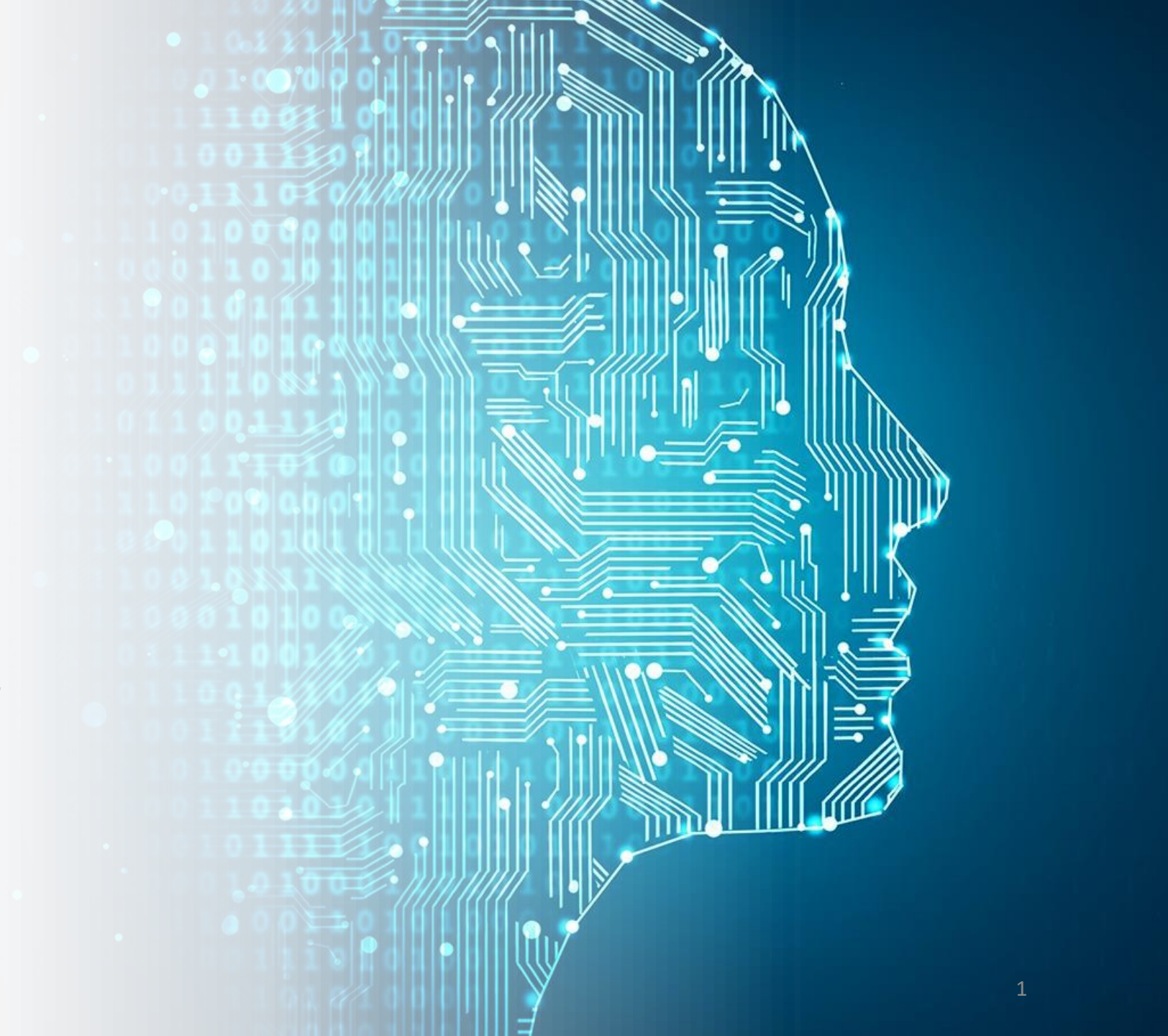
Topic Modeling Of Arabic Articles

Roy El Tenn

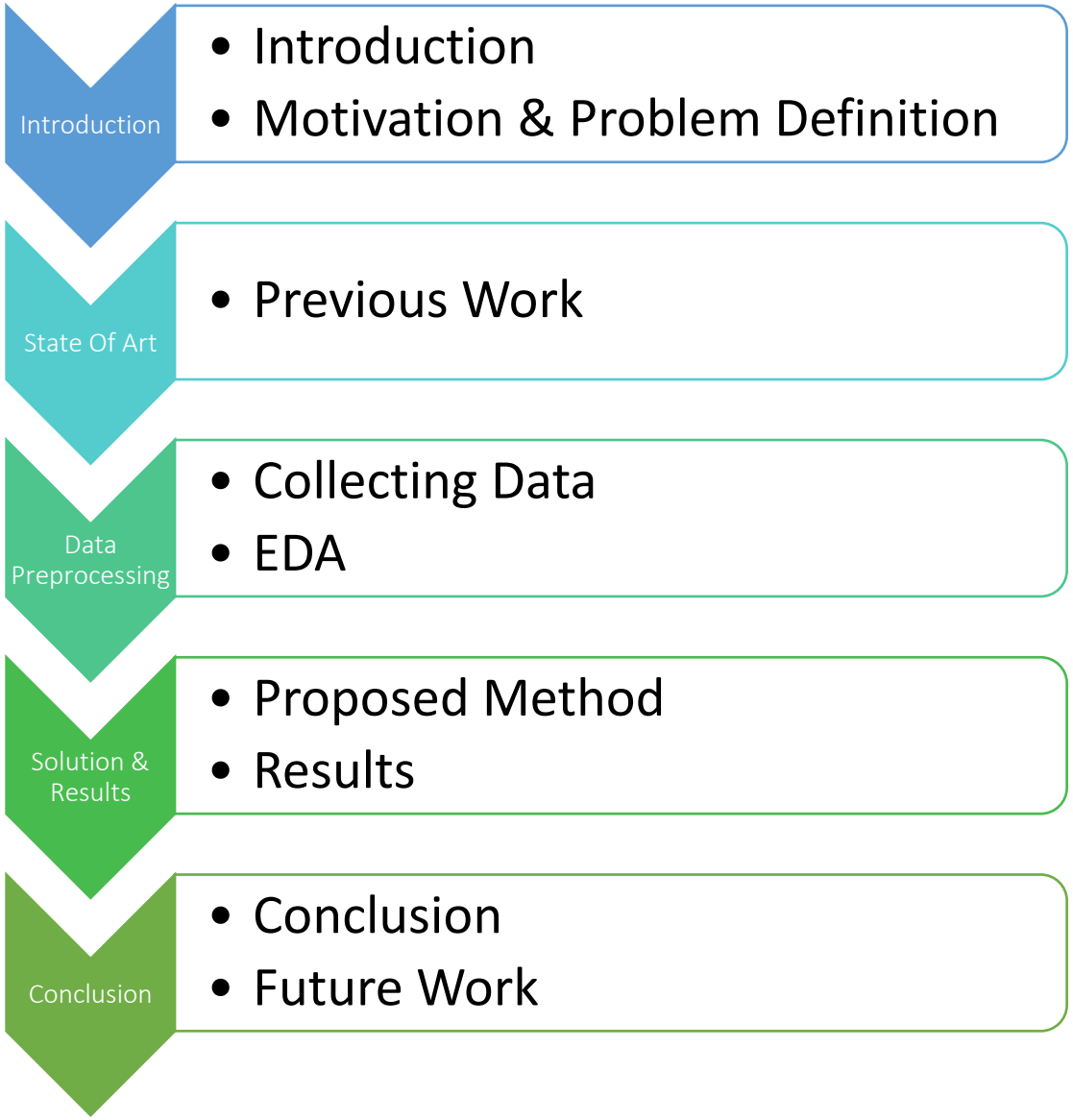
Ahmad Al Sharbaji

Capstone Project

Machine Learning Certification



OUTLINE


Introduction
• Introduction
• Motivation & Problem Definition
State Of Art
• Previous Work
Data Preprocessing
• Collecting Data
• EDA
Solution & Results
• Proposed Method
• Results
Conclusion
• Conclusion
• Future Work

Introduction

- Introduction
- Motivation & Problem Definition

Introduction

- Aims to build a model that can detect topics of texts in the Arabic language.
- Analyze the document data and predict one or more labels.

MLC: Multi-Label Classification is the problem of classifying instances into one or more classes where one or multiple labels can be assigned for each instance

	Belarus In Turmoil	COVID-19	How the Karabakh conflict is transforming regional politics	Myanmar Coup	Myanmar Election 2020	NewsFrames	آداب	أخبار جيدة	أخبار عاجلة	أديان	أصالة	أفكار	أفلام	احتجاج	الأصوات الصاعدة	الأعراق والأجناس	الألعاب الأولمبية	الإعلام والصحافة	الاقتصاد والأعمال	الجسر	الدعم الإنساني	النساء والنوع	النشاط الرقمي	الهجرة والنزوح	انتخابات
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1
3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
4965	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4966	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4967	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0

Problem Definition

Detect topics of text in the Arabic Language

قدمت مدونة شينجيانج أقصى غرب الصين مدينة كاراماي في شينجيانج، على أنها أغنى مدينة في الصين لعام 2012 وفقاً لأحدث بحث. يعود الفضل في ثروة المدينة المنشأة حديثاً إلى البترول

- صحافة المواطن
- تطوير

State Of Art

- Previous Work

- State of Art

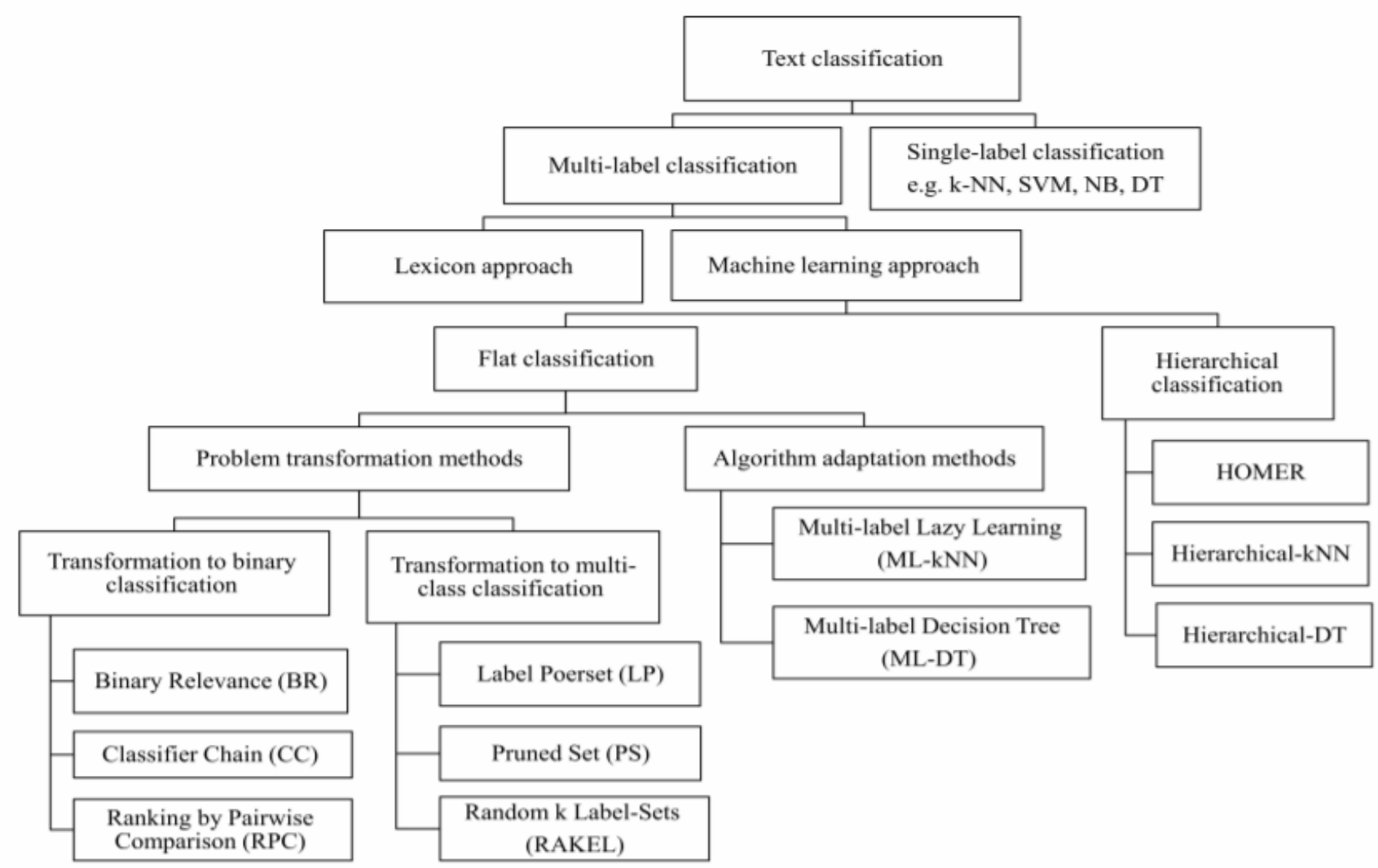


Fig. 1. Text Classification Taxonomy

Previous Solutions

- Few studies have investigated multi-label text classification for the Arabic language.
- Most of these studies have focused mainly on flat classification and have neglected the hierarchical structure.

ahmed et al : transform MLC into single label classification using MEKA tool to implement LP, BR approach.

Using SVM as a base classifier with the LP method achieved the best ML-accuracy with 71%.

Shehab et al. : Three multi-label classifiers were adapted to deal with MLC problems: random forest (RF), DT, and k-NN with k = 5 (5-NN).

The DT classifier achieved a better performance than the RF and 5-NN classifiers.

Previous Solutions

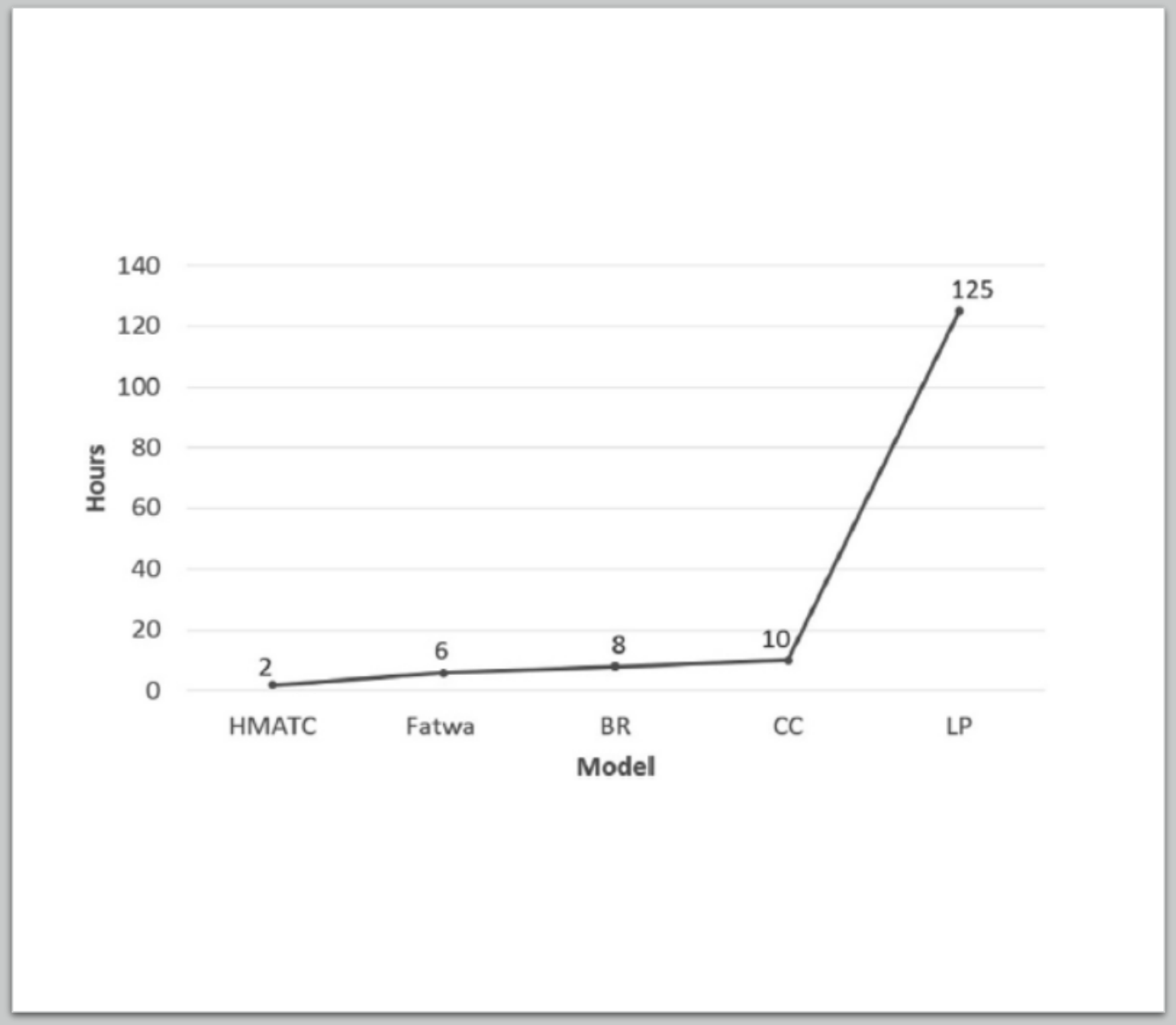
Zayed et al. hierarchical multi-label classification model to address HMC problems in the Arabic language which used the HOMER algorithm.

Achieve more effective predictive performance compared to the BR-NB classifier, which simply classified each label independently.

HMATC model was implemented based on the HOMER algorithm, which was optimized by employing LP-SVM classifier and balanced k-means algorithm with eight clusters in order to improve hierarchical classification outcomes.

HMATC model outperformed all the evaluated models in terms of computational cost.

- **Previous Solutions**



Data Preprocessing

- Collecting Data
- EDA

Data Preprocessing

- Data Provided by TARJAMA team
- Text & Label Separated

- Merge the Data
- Removed the unused columns
- Drop the Null Text rows

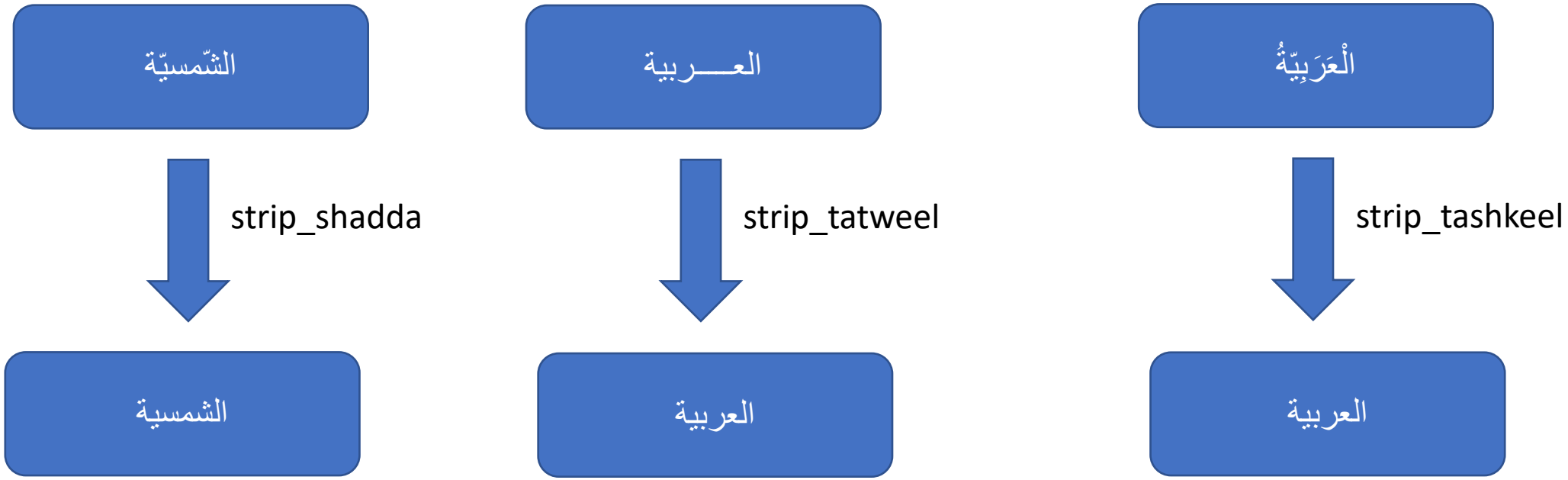
1	Belarus In Turmoil	COVID-19	How the Karabakh conflict is transforming regional politics	Myanmar Coup	Myanmar Election 2020	NewsFrames	آداب	أخبار جيدة	أخبار عاجلة	أديان	أصالة	أفكار	أفلام	احتجاج	الأصوات الصاخبة	الأعراق والأجناس	الألعاب الأولمبية	الإعلام والصحافة	الاقتصاد والأعمال	الجسر	الدعم الإنساني	النساء والنوع	النشاط الرقمي	الهجرة والتزوج	انتخابات	بود كاست الأصوات العالمية
لخصت الفنانة الجامايكية جودي آن ماكميلان سيرة ...	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
في الأسبوع الماضي، ثار الجمهور نتيجة لتصريحات...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
جميع الروابط] تؤدي للفرنسية ما لم ينص على خلاف...	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0
مواطنو أمريكا اللاتينية يحبون السفر. يمكن أن ي...	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
اليوم مر عام منذ أن تم اعتقال وسجن باسل خرطبيل...	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Data Preprocessing

Some function :

- strip_tashkeel: bool = True
- strip_tatweel: bool = True
- insert_white_spaces
- strip_ shadda : bool = True

- Used ArabertPreprocessor function from arabert.preprocess



- Increased the amount of data by using ML SMOTE

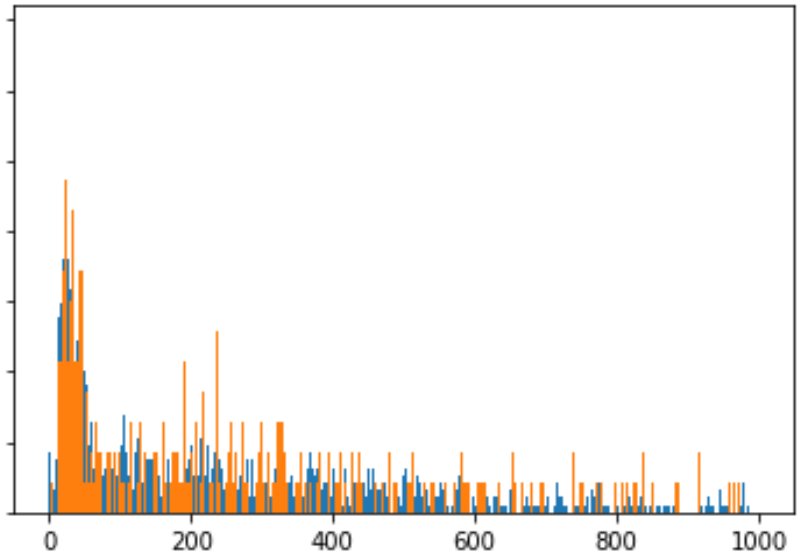
Data Preprocessing

- Split our data into train 80 % & test 20 %

```
train_AJGT, test_AJGT = train_test_split(df_AJGT, test_size=0.2,random_state=42)
```

```
!mkdir data
train_AJGT.to_csv("data/train.csv",index=False,columns=train_AJGT.columns,sep='\t',header=True)
test_AJGT.to_csv("data/test.csv",index=False,columns=test_AJGT.columns,sep='\t',header=True)
```

- Max Length of Training data is 3274
- Max Length of Testing Data is 2930

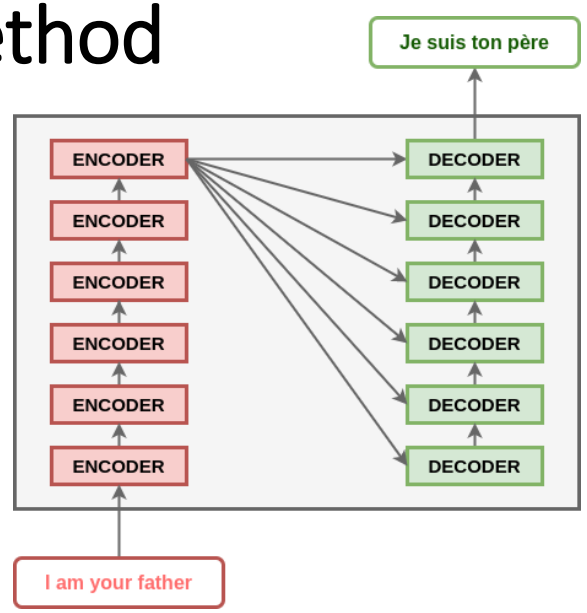


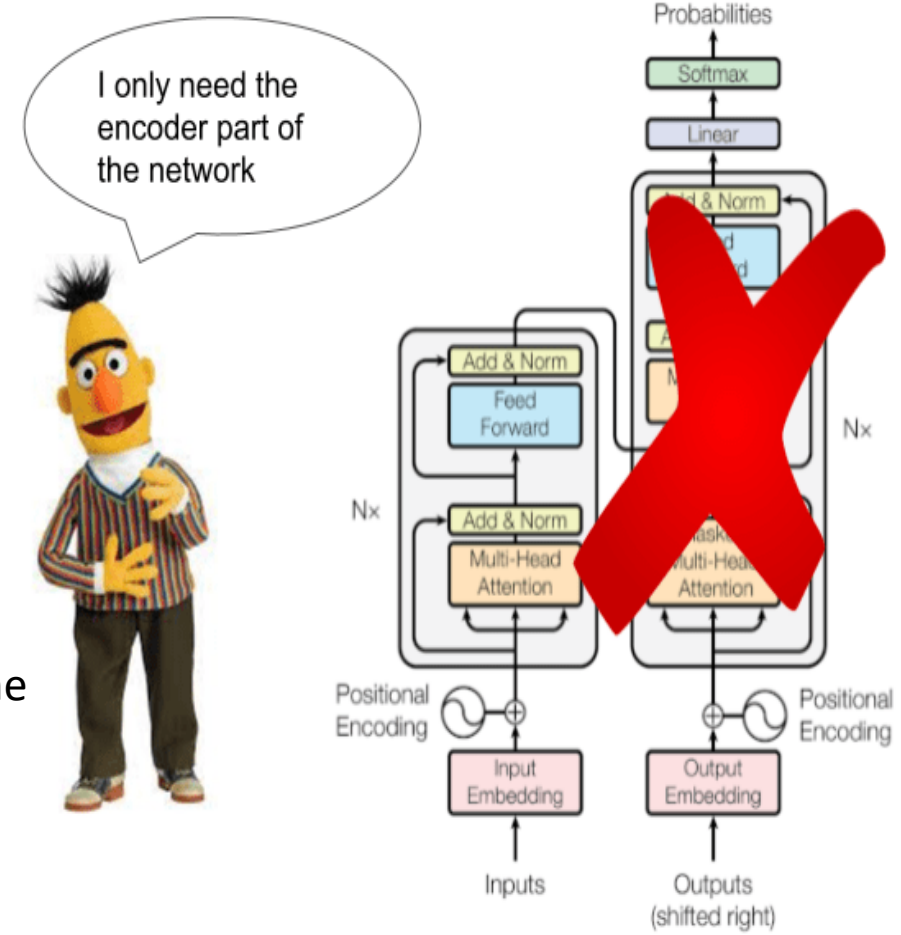
	support	itemsets
0	0.186154	(احتجاج)
1	0.125176	(الإعلام والصحافة)
2	0.197827	(النشاط الرقمي)
3	0.125176	(تقنية)
4	0.169853	(حروب ونزاعات)
5	0.184142	(حرية التعبير)
6	0.276313	(حقوق الإنسان)
7	0.129000	(حكم)
8	0.290199	(سياسة)
9	0.101429	(شباب)
10	0.750453	(صحافة المواطن)
11	0.102838	(علاقات دولية)
12	0.147515	(فنون وثقافة)
13	0.118334	(سياسة, احتجاج)
14	0.163011	(احتجاج, صحافة المواطن)
15	0.152948	(النشاط الرقمي, صحافة المواطن)
16	0.130409	(صحافة المواطن, حروب ونزاعات)
17	0.112497	(حقوق الإنسان, حرية التعبير)
18	0.146911	(صحافة المواطن, حرية التعبير)
19	0.109076	(سياسة, حقوق الإنسان)
20	0.232240	(حقوق الإنسان, صحافة المواطن)
21	0.100020	(صحافة المواطن, حكم)

Solution & Results

- Proposed Method
- Results

Proposed Method





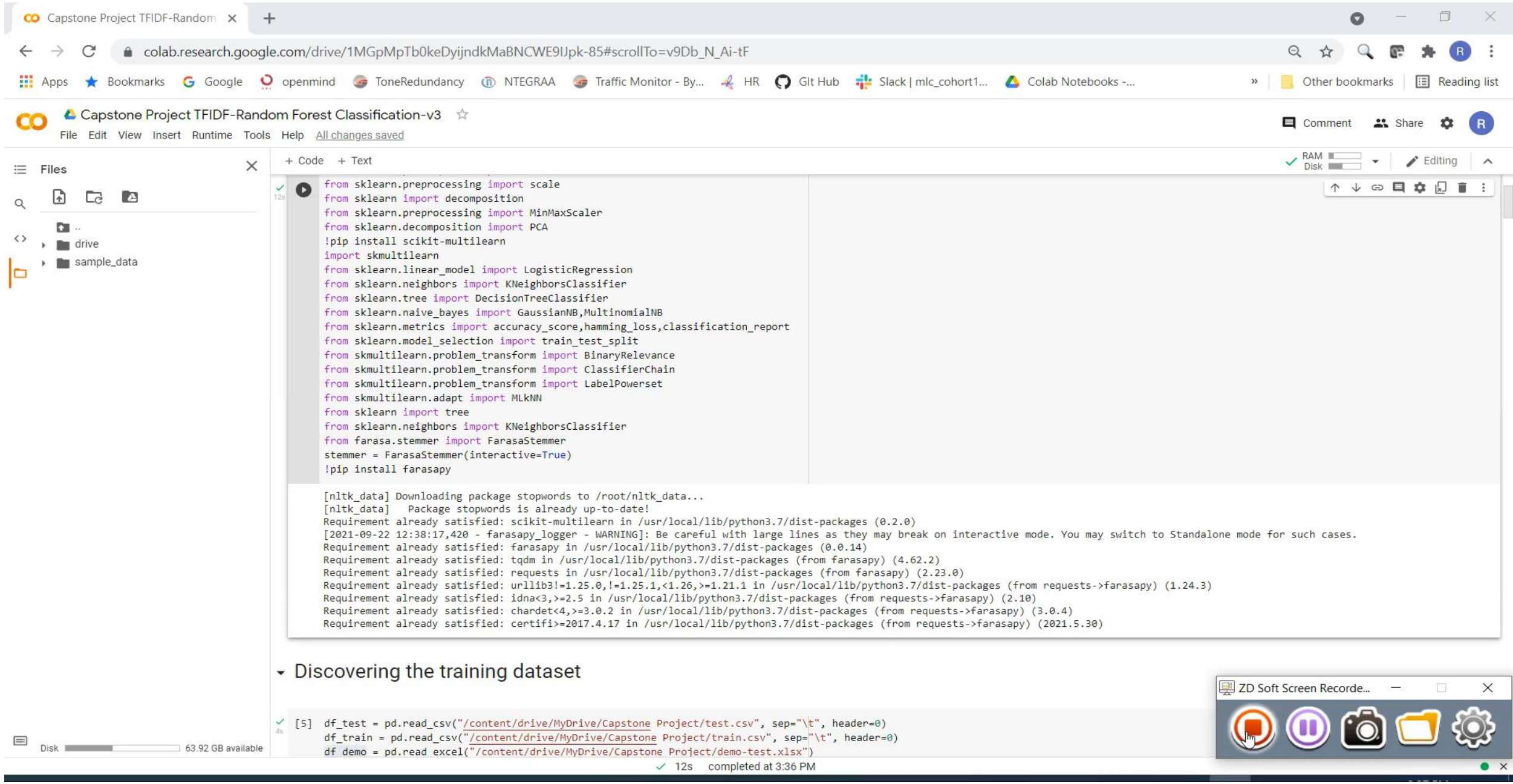
- Transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side.

- BERT is basically an Encoder stack of transformer architecture.

- AraBERT is an Arabic pretrained language model based on Google's BERT architecture

Results

	TFIDF	
	Random Forest	SVM
Exact Match Ratio	0.75	0.24
F1-Score Weighted	0.87	0.64
F1-Score Samples	0.83	0.63
Hamming Loss	0.03	0.08


Capstone Project TFIDF-Random Forest Classification-v3
colab.research.google.com/drive/1MGpMpTb0keDyijndkMaBNCWE9IJpk-85#scrollTo=v9Db_N_Ai-tF
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
drive
sample_data
+ Code + Text
RAM
Disk
Editing
from sklearn.preprocessing import scale
from sklearn import decomposition
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
!pip install scikit-multilearn
import skmultilearn
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB,MultinomialNB
from sklearn.metrics import accuracy_score,hamming_loss,classification_report
from sklearn.model_selection import train_test_split
from skmultilearn.problem_transform import BinaryRelevance
from skmultilearn.problem_transform import ClassifierChain
from skmultilearn.problem_transform import LabelPowerset
from skmultilearn.adapt import MLkNN
from sklearn import tree
from sklearn.neighbors import KNeighborsClassifier
from farasa.stemmer import FarasaStemmer
stemmer = FarasaStemmer(interactive=True)
!pip install farasapy
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
Requirement already satisfied: scikit-multilearn in /usr/local/lib/python3.7/dist-packages (0.2.0)
[2021-09-22 12:38:17,420 - farasapy_logger - WARNING]: Be careful with large lines as they may break on interactive mode. You may switch to Standalone mode for such cases.
Requirement already satisfied: farasapy in /usr/local/lib/python3.7/dist-packages (0.0.14)
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from farasapy) (4.62.2)
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from farasapy) (2.23.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requests->farasapy) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests->farasapy) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from requests->farasapy) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests->farasapy) (2021.5.30)
Discovering the training dataset
[5] df_test = pd.read_csv("/content/drive/MyDrive/Capstone Project/test.csv", sep="\t", header=0)
df_train = pd.read_csv("/content/drive/MyDrive/Capstone Project/train.csv", sep="\t", header=0)
df_demo = pd.read_excel("/content/drive/MyDrive/Capstone Project/demo-test.xlsx")
Disk 63.92 GB available
12s completed at 3:36 PM
ZD Soft Screen Recorde...

Conclusion

- Conclusion
- Future Work

Future Work

- Try To implement hierarchical Solution.
- Results based on Arabert.
- Tunning the model.

ANY Questions?

Thank You…