VIETNAMESE TEXT CLASSIFICATION

Candidate: Nguyễn Hoàng Dũng

TABLE OF CONTENTS

TABLE OF CONTENTS

01 INTRODUCTION

What is text classification? Its approaches?

02 ML PROCESS

From data understanding, preprocessing, and modelling

03 EVALUATION

Is the work do well? What metrics to be used?

04 CONSIDERATION

Is this machine learning algorithm effective for this data?

TEXT CLASSIFICATION

Text classification is a task that assigns a set of predefined categories to open-ended text.

Due to the categories, we can divide text classification into smaller subtasks.

Sub-task	**Categories**
Spam detection	[‘Spam’, ‘Not Spam’]
Intent detection	[‘Has intent’, ‘No intent’]
Sentiment analysis	[‘Like’, ‘Dislike’, ‘Neutral’]
Topic labeling	[‘Sports’, ‘Business’, ‘Travel’, ‘Culture’, ‘Tech’, ...]

TEXT CLASSIFICATION APPROACHES

RULE-BASED

Use a set of handcrafted linguistic rules.

ML-BASED

Let the machine learns to make classifications based on labeled samples.

HYBRID

Combine ML-based with rule-based to improve the results

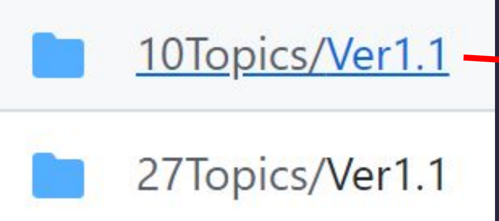
WHAT MY ASSIGNMENT APPLIED?

SUBTASK	APPROACH	LANGUAGE
Spam detection	Rule-based	Vietnamese
Sentiment analysis	ML-based	English
Topic modeling	Hybrid	Other

DATASET

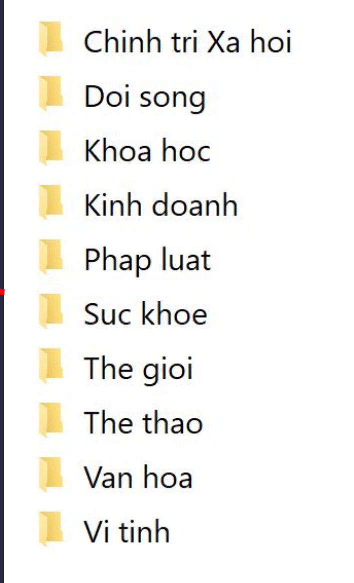
From the paper: **A Comparative Study on Vietnamese Text Classification Methods**

Link: https://github.com/duyvuleo/VNTC



10Topics/Ver1.1

27Topics/Ver1.1



Chinh tri Xa hoi

Doi song

Khoa hoc

Kinh doanh

Phap luat

Suc khoe

The gioi

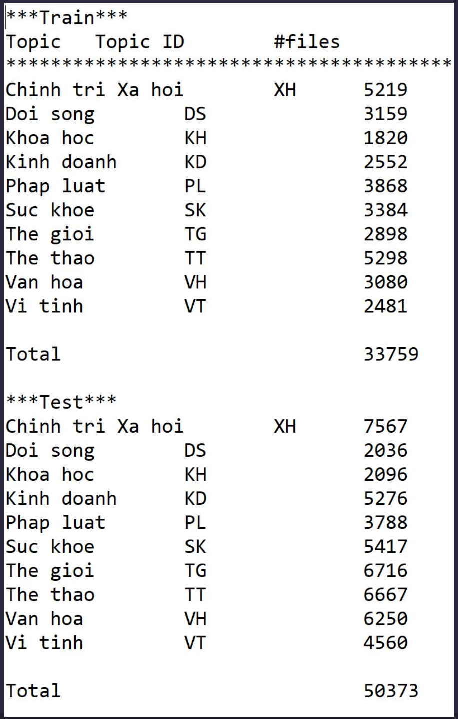
The thao

Van hoa

Vi tinh

2 types of labeling

DATASET



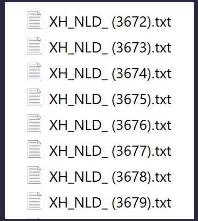
```
***Train***
Topic   Topic ID        #files
****************************************
Chinh tri Xa hoi        XH      5219
Doi song        DS              3159
Khoa hoc        KH              1820
Kinh doanh      KD              2552
Phap luat       PL              3868
Suc khoe        SK              3384
The gioi        TG              2898
The thao        TT              5298
Van hoa         VH              3080
Vi tinh         VT              2481

Total                           33759

***Test***
Chinh tri Xa hoi        XH      7567
Doi song        DS              2036
Khoa hoc        KH              2096
Kinh doanh      KD              5276
Phap luat       PL              3788
Suc khoe        SK              5417
The gioi        TG              6716
The thao        TT              6667
Van hoa         VH              6250
Vi tinh         VT              4560

Total                           50373
```

STATISTICS

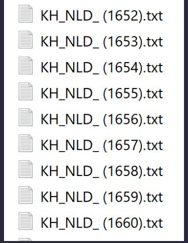
- Training set: **33579 samples**
- Test set: **50373 samples**

DATA PREPARATION

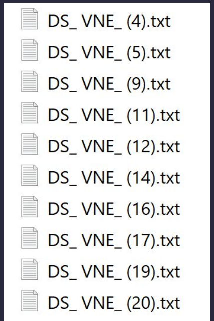
Converting all the files (in .txt) to 2 .csv files.
(Code shown in: csv_generation.ipynb)



XH_NLD_ (3672).txt
XH_NLD_ (3673).txt
XH_NLD_ (3674).txt
XH_NLD_ (3675).txt
XH_NLD_ (3676).txt
XH_NLD_ (3677).txt
XH_NLD_ (3678).txt
XH_NLD_ (3679).txt



KH_NLD_ (1652).txt
KH_NLD_ (1653).txt
KH_NLD_ (1654).txt
KH_NLD_ (1655).txt
KH_NLD_ (1656).txt
KH_NLD_ (1657).txt
KH_NLD_ (1658).txt
KH_NLD_ (1659).txt
KH_NLD_ (1660).txt



DS_ VNE_ (4).txt
DS_ VNE_ (5).txt
DS_ VNE_ (9).txt
DS_ VNE_ (11).txt
DS_ VNE_ (12).txt
DS_ VNE_ (14).txt
DS_ VNE_ (16).txt
DS_ VNE_ (17).txt
DS_ VNE_ (19).txt
DS_ VNE_ (20).txt

DATA PREPROCESSING

TEXT NORMALIZATION

WORD SEGMENTATION & TOKENIZATION

ID CONVERT & SEQUENCE PADDING

DATA PREPROCESSING

WORD SEGMENTATION

Perform stripping, lowercase and remove unwanted symbols.

“ Thành lập dự án POLICY phòng chống HIV/AIDS ở VN (NLĐ)- Quỹ hỗ trợ khẩn cấp về AIDS của Hoa Kỳ vừa thành lập dự án POLICY tại VN... “

→

“thành lập dự án policy phòng chống hiv aids ở vn nlđ quỹ hỗ trợ khẩn cấp về aids của hoa kỳ vừa thành lập dự án policy tại vn“

DATA PREPROCESSING

WORD SEGMENTATION & TOKENIZATION

Text segmentation is the process of dividing written text into meaningful units.

Why we need to do it?
In Vietnamese, a word may have it meaning changes when follow by another word.

E.g.,
Hoa = flowers, Kỳ = strange, Hoa Kỳ = U.S

DATA PREPROCESSING

WORD SEGMENTATION & TOKENIZATION

“thành lập dự án policy phòng chống hiv aids ở vn nlđ quỹ hỗ trợ khẩn cấp về aids của hoa kỳ vừa thành lập dự án policy tại vn“

→

['thành_lập', 'dự_án', 'policy', 'phòng_chống', 'hiv', 'aids', 'ở', 'vn', 'nlđ', 'quỹ', 'hỗ_trợ', 'khẩn_cấp', 'về', 'aids', 'của', 'hoa_kỳ', 'vừa', 'thành_lập', 'dự_án', 'policy', 'tại', 'vn']

DATA PREPROCESSING

ID CONVERT & SEQUENCE PADDING

['thành_lập', 'dự_án', 'policy', 'phòng_chống', 'hiv', 'aids', 'ở', 'vn', 'nlđ', 'quỹ', 'hỗ_trợ', 'khẩn_cấp', 'về', 'aids', 'của', 'hoa_kỳ', 'vừa', 'thành_lập', 'dự_án', 'policy', 'tại', 'vn']

[0, 763, 169, 3, 2137, 3, 3, 25, 33756, 3, 1425, 291, 2498, 28, 3, 7, 3, 164, 763, 169, 3, 35, 33756]

[0, 763, 169, 3, 2137, 3, 3, 25, 33756, 3, 1425, 291, 2498, 28, 3, 7, 3, 164, 763, 169, 3, 35, 33756, 0, 0, 0, 0, 0, 0, 0, 0, 0,...]

MODELING

ML-BASED MODEL

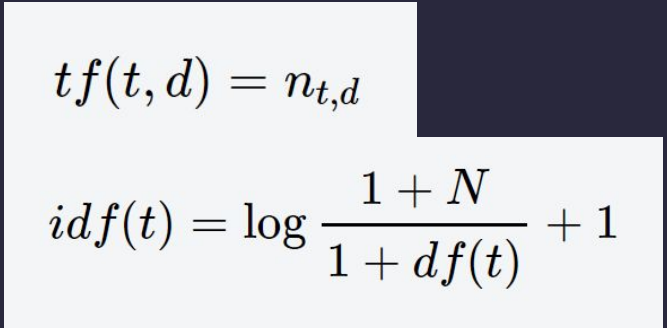
Vectorization: TF-IDF (Term Frequency - Inverse Document Frequency)
Classifier: SVM (Support Vector Machine)

DL-BASED MODEL

Vectorization: One-hot word representation
Classifier: CNN (Convolutional Neural Network)

MODELING

TF-IDF (scikit-learn version):

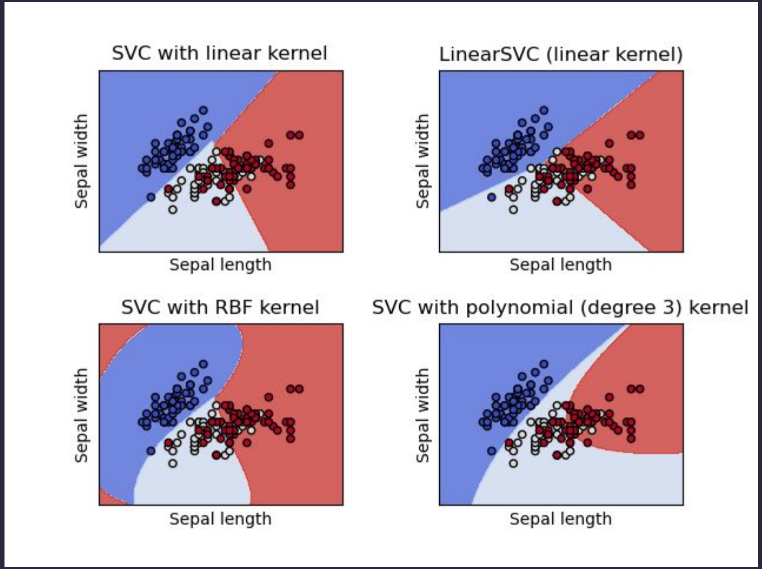


$$tf(t,d) = n_{t,d}$$

$$idf(t) = \log \frac{1+N}{1+df(t)} + 1$$

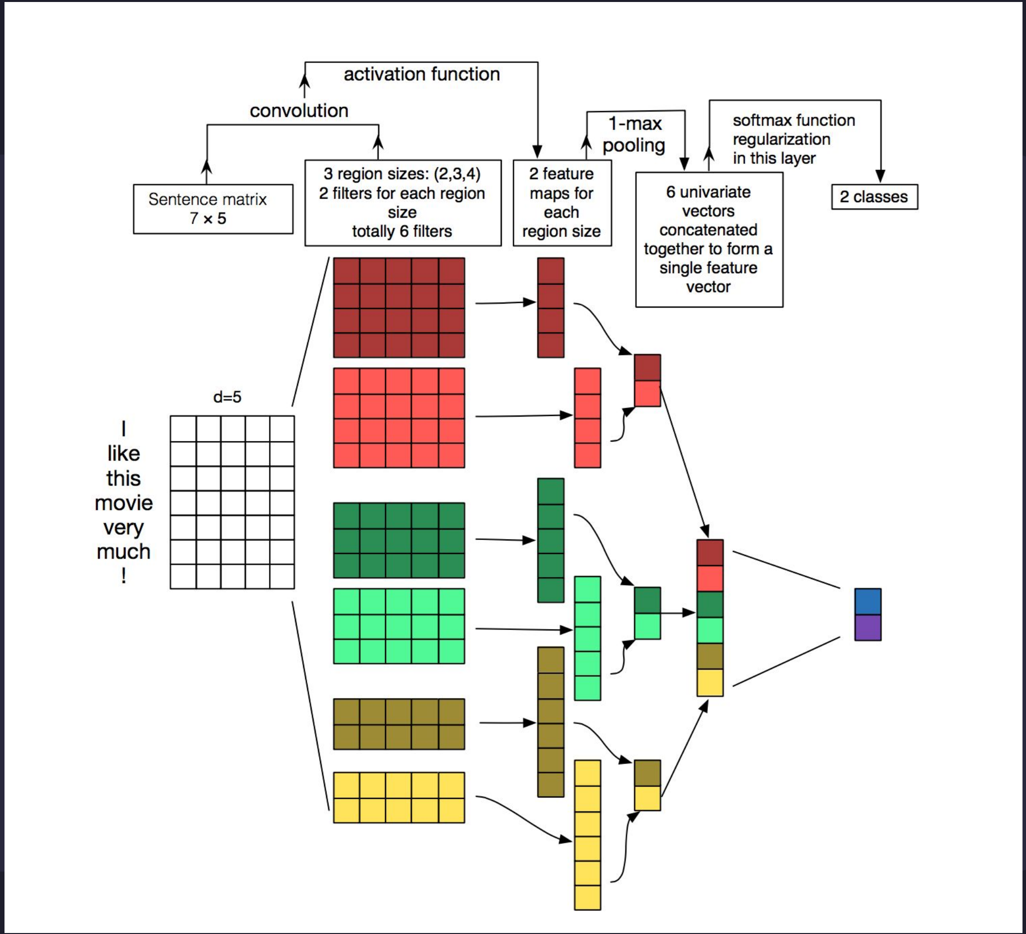
MODELING

SVM (support Vector Machine):



MODELING

Convolutional Neural Network (CNN)


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

MODELING

1. Take IDs as inputs, Tensorflow will convert every tokens into a one-hot vector.
2. Have an Embedding layer to reduce the dimension.
3. Apply Conv1D layers with different filter sizes (2, 3 and 4), use ReLU activation functions, followed by GlobalMaxPool1D layers.
4. Concatenate the output and pass through fully-connected layers (ReLU activation functions) applied dropout regularization.
5. The final layer is a softmax layer.

MODELING

OPTIMIZER	ADAM
LOSS	CATEGORICAL CROSS-ENTROPY
EPOCH	20
BATCH SIZE	32

EVALUATION METRICS

ACCURACY

Number of correct predictions

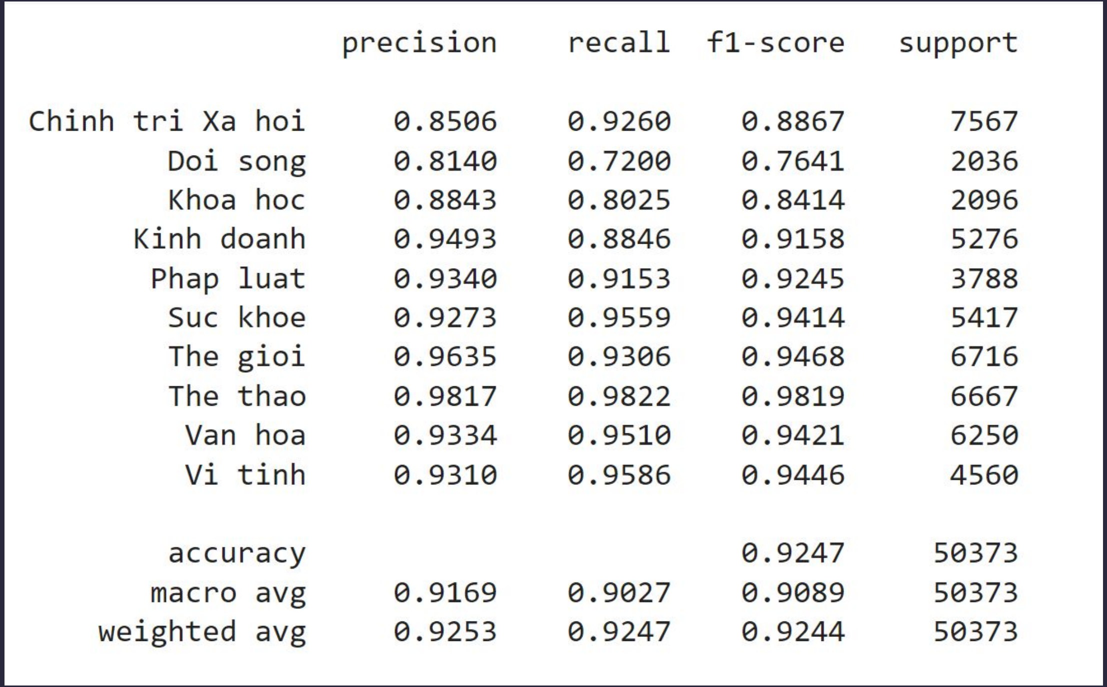
PRECISION

Number of positive class predictions that actually belong to the positive class over the test set.

RECALL

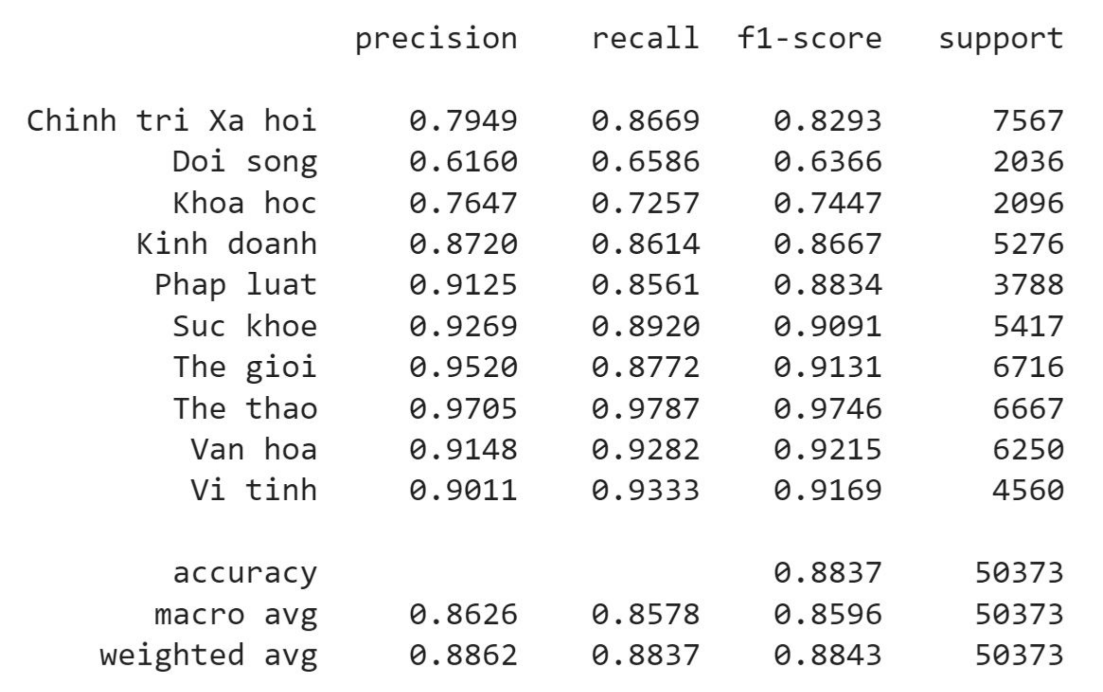
Number of positive class predictions made out of all positive samples.

EVALUATION (SVM)



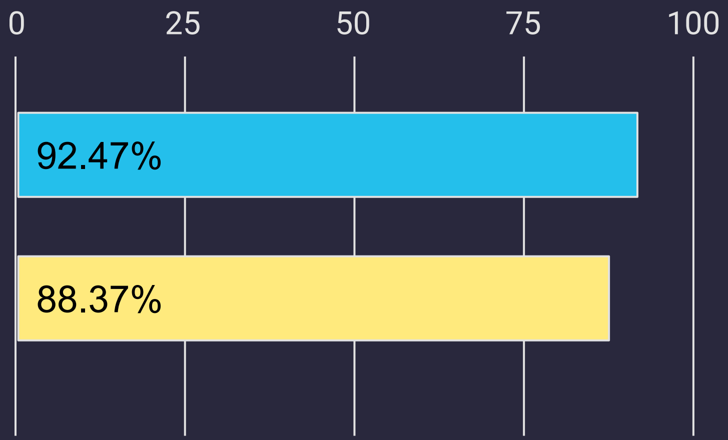
```
                   precision    recall  f1-score   support

Chinh tri Xa hoi      0.8506    0.9260    0.8867      7567
        Doi song      0.8140    0.7200    0.7641      2036
        Khoa hoc      0.8843    0.8025    0.8414      2096
      Kinh doanh      0.9493    0.8846    0.9158      5276
       Phap luat      0.9340    0.9153    0.9245      3788
        Suc khoe      0.9273    0.9559    0.9414      5417
        The gioi      0.9635    0.9306    0.9468      6716
        The thao      0.9817    0.9822    0.9819      6667
         Van hoa      0.9334    0.9510    0.9421      6250
         Vi tinh      0.9310    0.9586    0.9446      4560

        accuracy                          0.9247     50373
       macro avg      0.9169    0.9027    0.9089     50373
    weighted avg      0.9253    0.9247    0.9244     50373
```

EVALUATION (CNN)



```
                  precision    recall  f1-score   support

Chinh tri Xa hoi     0.7949    0.8669    0.8293      7567
       Doi song     0.6160    0.6586    0.6366      2036
        Khoa hoc     0.7647    0.7257    0.7447      2096
     Kinh doanh     0.8720    0.8614    0.8667      5276
      Phap luat     0.9125    0.8561    0.8834      3788
        Suc khoe     0.9269    0.8920    0.9091      5417
        The gioi     0.9520    0.8772    0.9131      6716
        The thao     0.9705    0.9787    0.9746      6667
         Van hoa     0.9148    0.9282    0.9215      6250
        Vi tinh     0.9011    0.9333    0.9169      4560

        accuracy                         0.8837     50373
       macro avg     0.8626    0.8578    0.8596     50373
    weighted avg     0.8862    0.8837    0.8843     50373
```

CONSIDERATION



Accuracy comparison of 2 models

● **Support Vector Machine (SVM)**

With TF-IDF vectorizer.

● **Convolutional Neural Network (CNN)**

With one-hot word representation.

THANK YOU FOR LISTENING!

Also thank our mentors for your valuable guidance in this course!