ETH zürich

Web Type Classification

Mid-Term

Jie Huang

Catalogue

- Introduction
- Dataset
- Model based on HTML content
- Model based on screenshots
- Experiments and Results
- Problem and Future Work

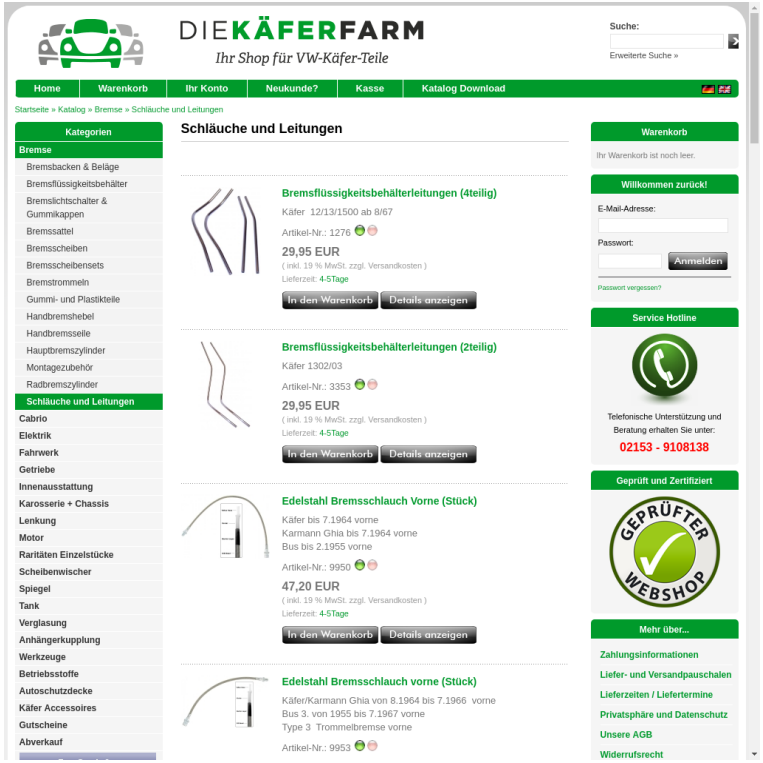
Catalogue

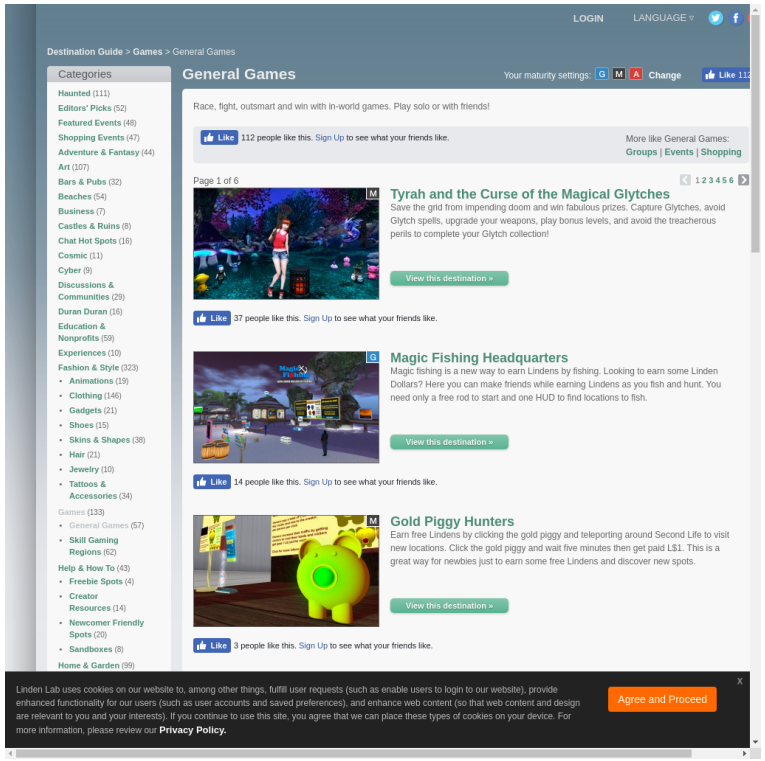
Introduction

- Web Type Classification:
  - Classify web pages via the screenshot and web source code
- Motivation:
  - Many tasks that process websites (e.g. boilerplate removal, summarization, link analysis, …) benefit from knowing what type of website they are dealing with
- Types:
  - Collection
  - Profile
  - Media Item
  - Discussion
  - Textual
  - Other

Collection

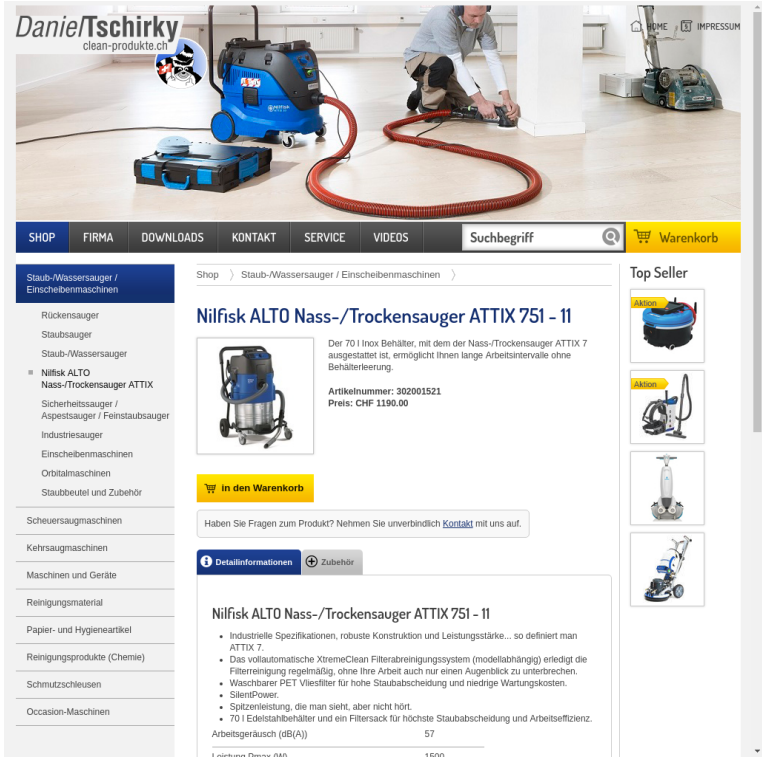
- Directories
- Search Results
- Homepages (if they are a collections of links)
- Lists of discussion threads

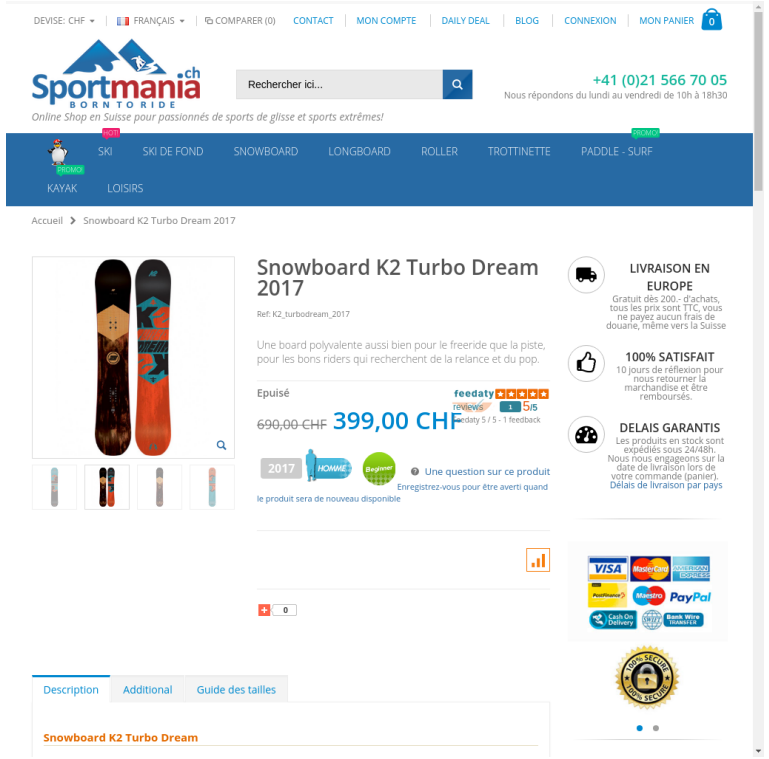




Profile

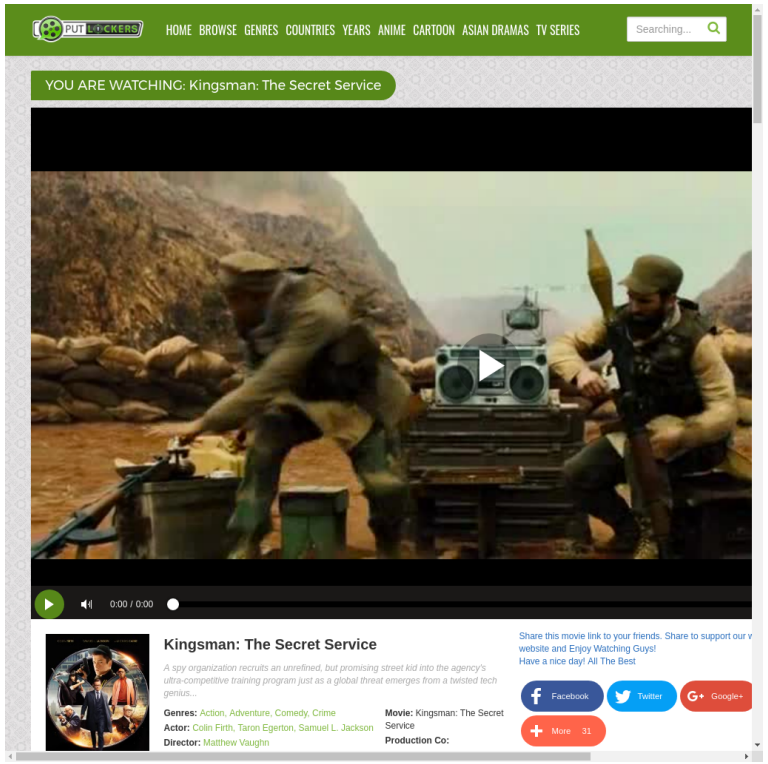
- Structured information about an entity:
  - Products
  - Persons

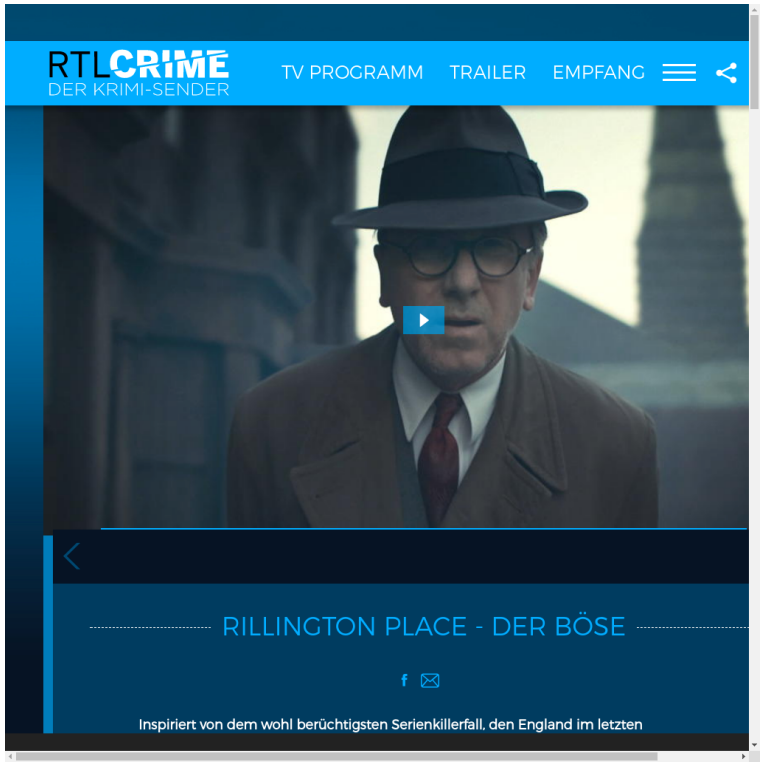




Media Item

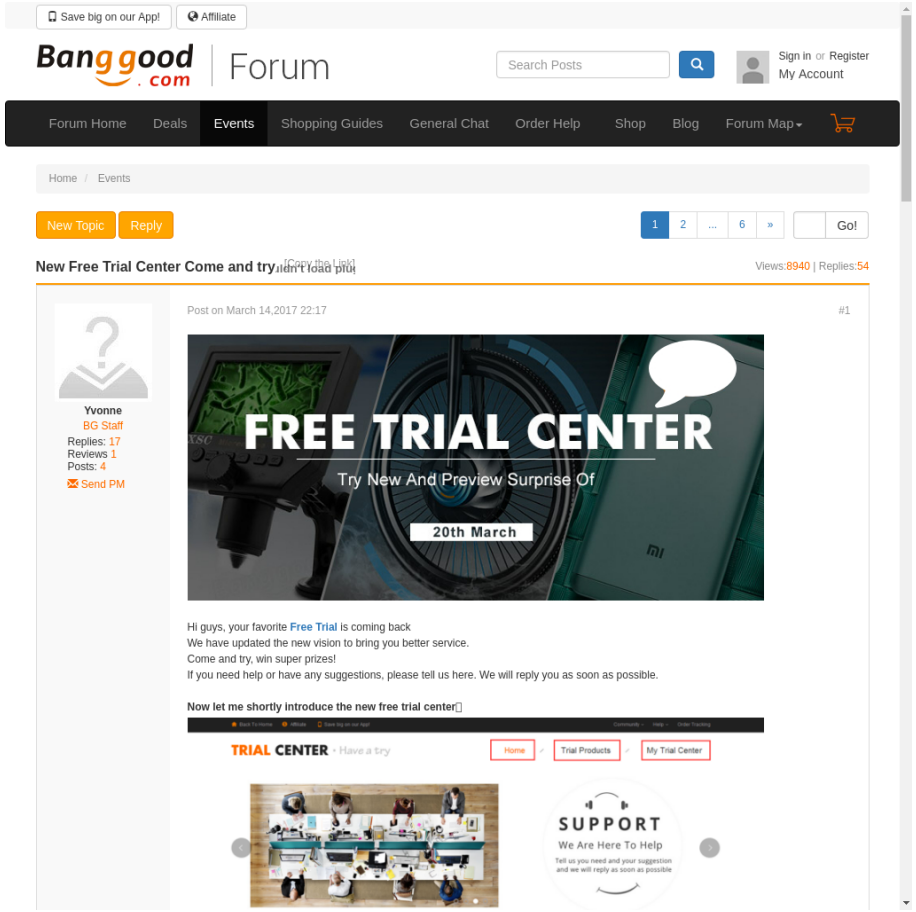
- Images
- Videos
- Songs





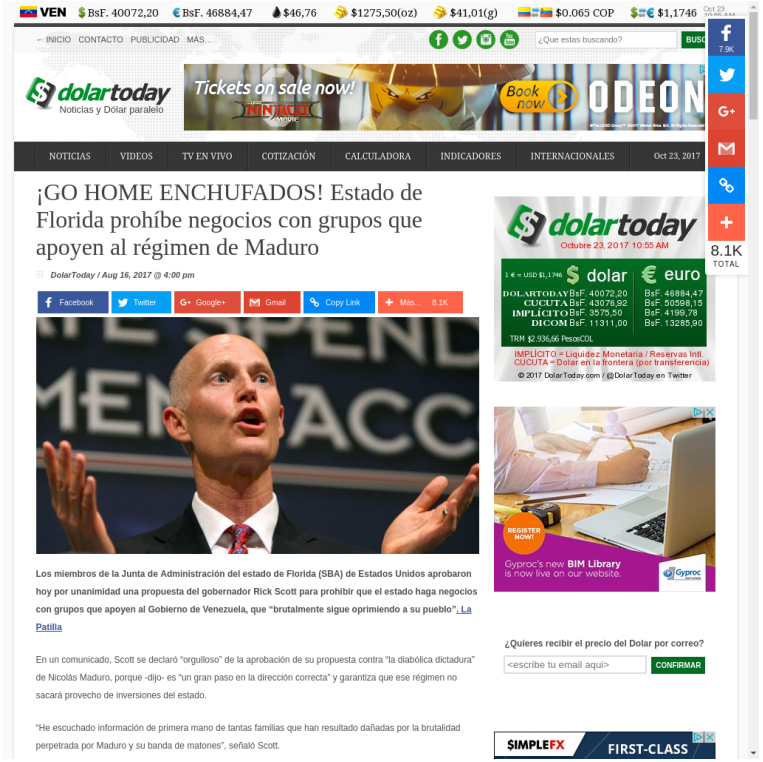
Discussion

- Threads
- Reviews
- Comments



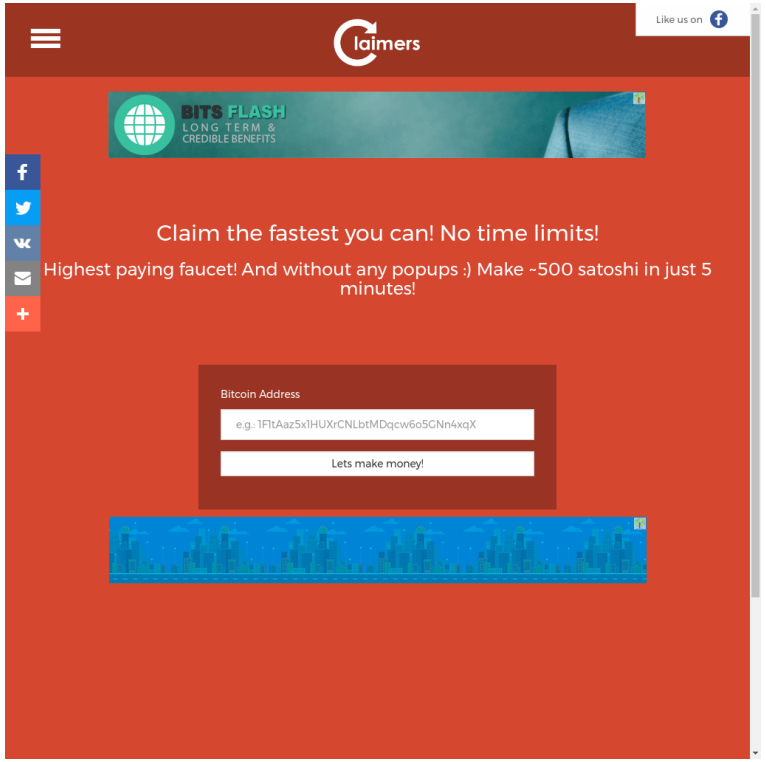
Textual

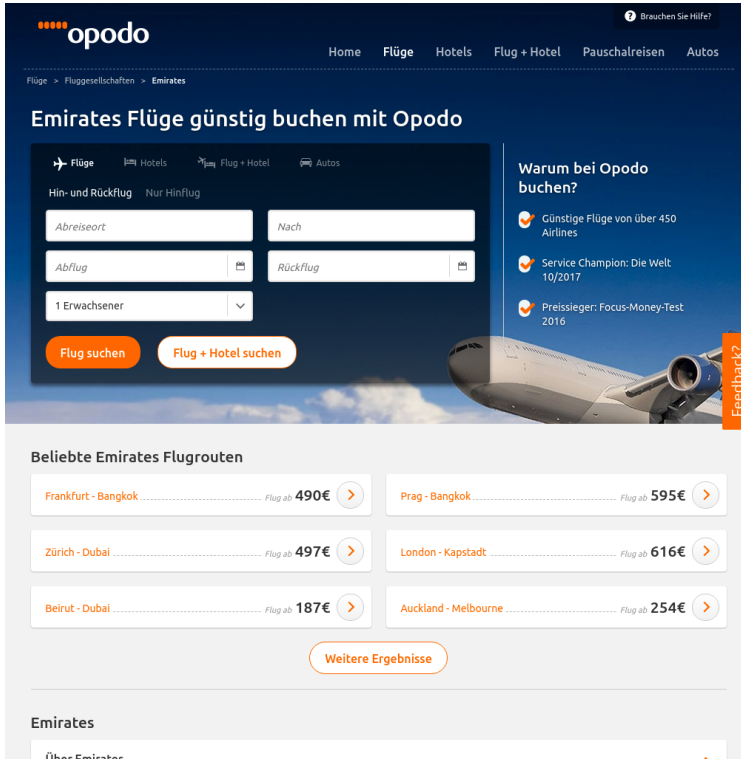
- Newspaper article
- Blog
- Generic information
- Static page

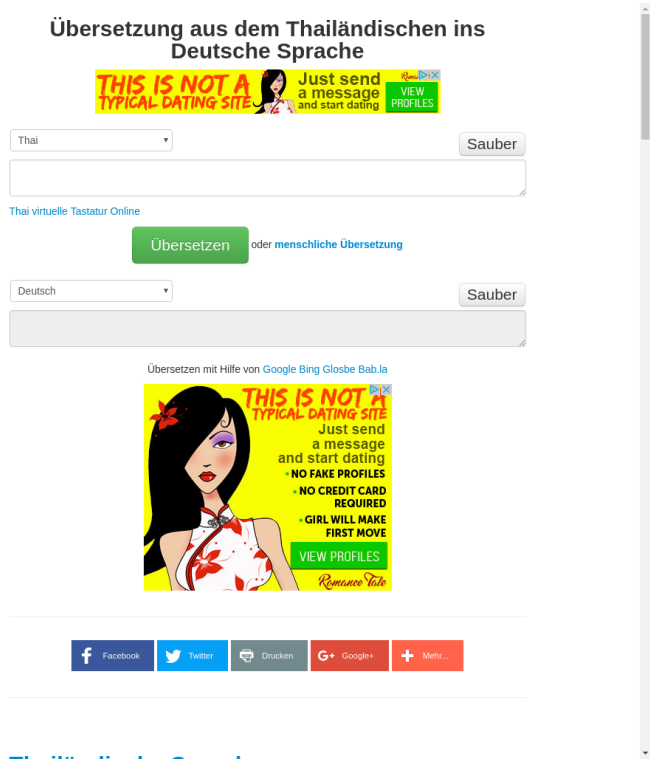


Other

- Contact forms
- Search pages (not results)
- Maps
- Pages with no clearly recognizable structure







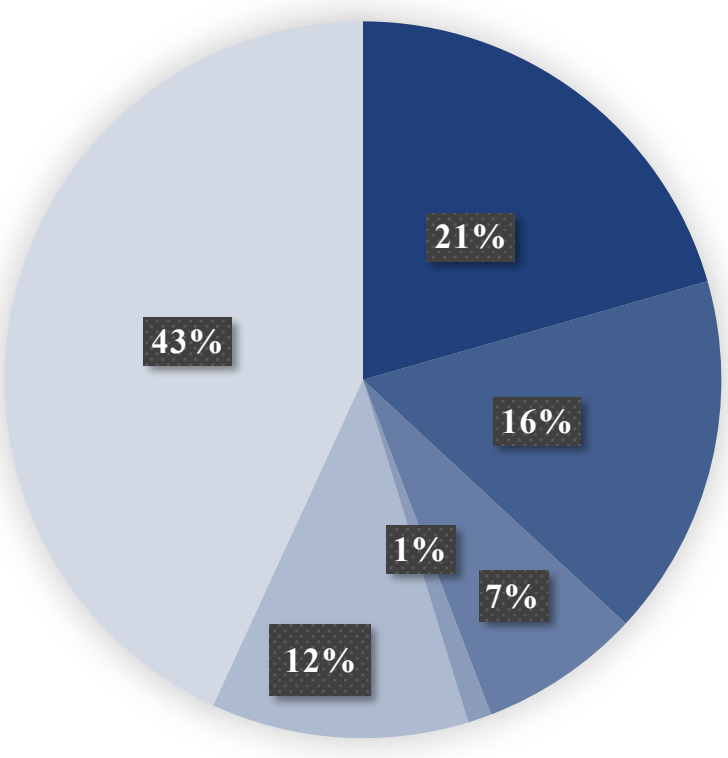
Catalogue

- Introduction
- Dataset
- Model based on HTML content
- Model based on screenshots
- Experiments and Results
- Problem and Future Work

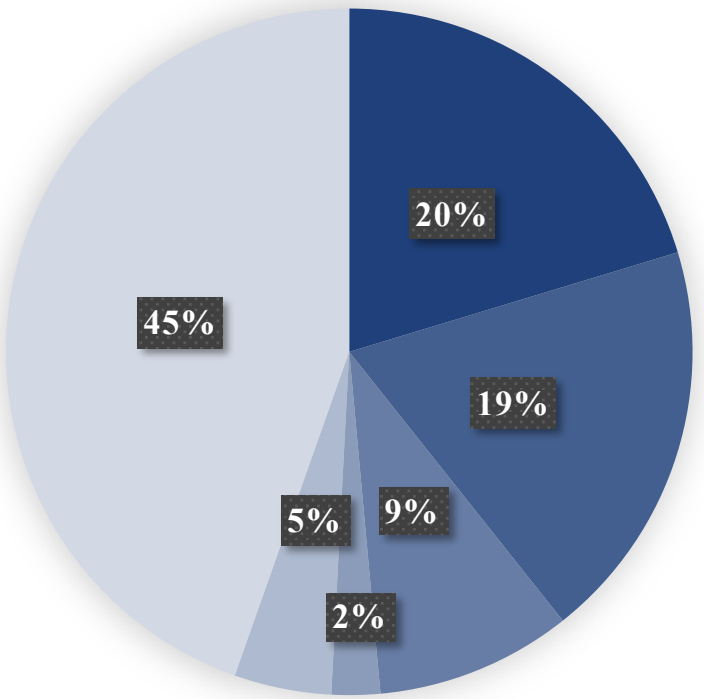
Dataset - Totally

Type	Size	Details
HTML	93207	HTML Pages
Screenshots	89434	1024*1024 RBG Image

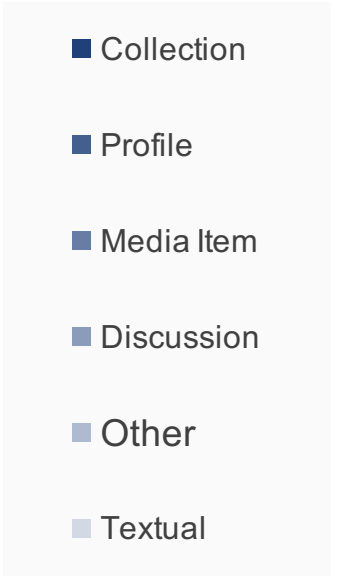
Dataset - Labeled

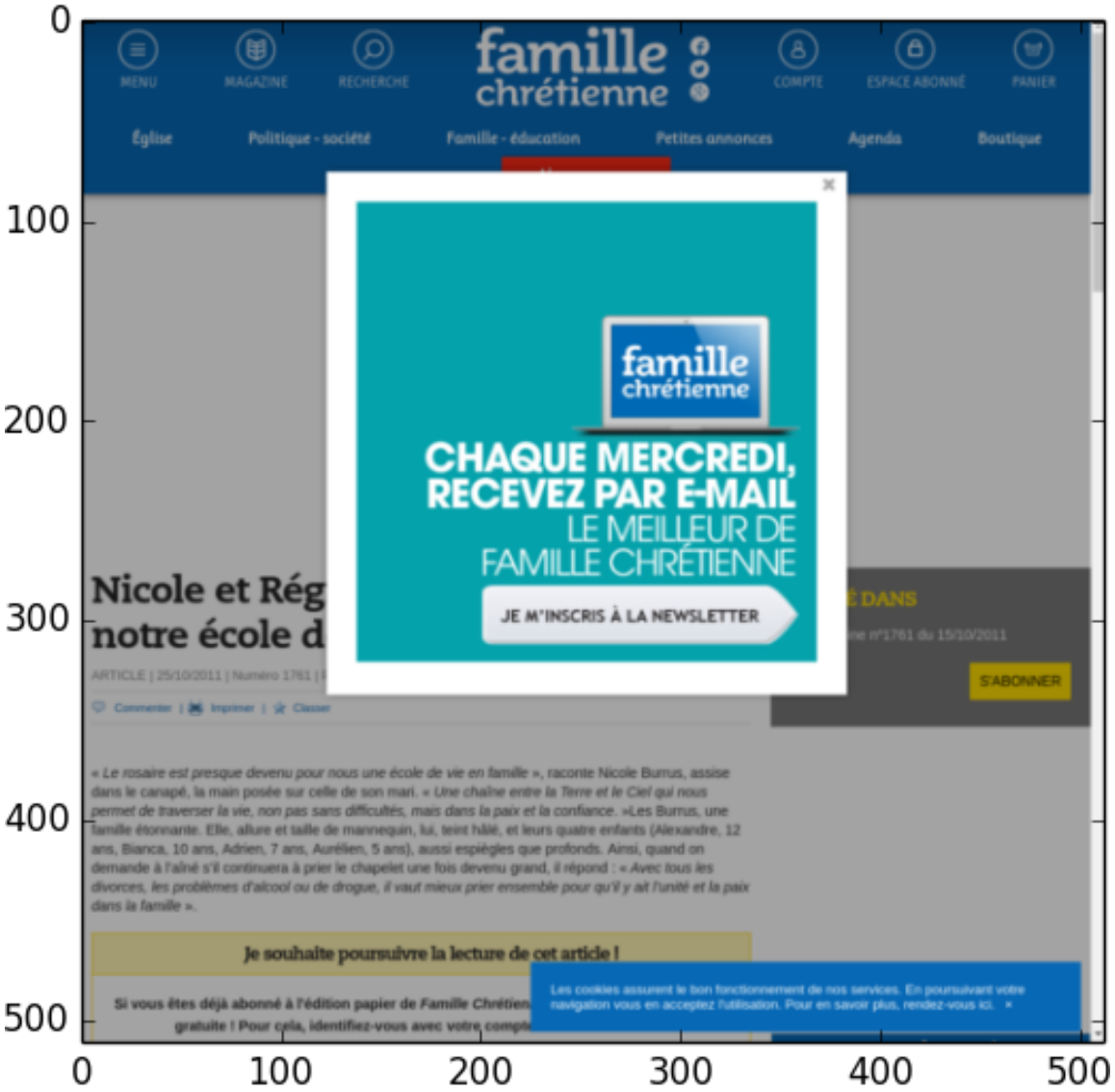


Labeled dataset – Totally
Size: 2791



Labeled dataset – Clean
Size: 1137




famille chrétienne
MENU
MAGAZINE
RECHERCHE
COMPTE
ESPACE ABONNÉ
PANIER
Église
Politique - société
Famille - éducation
Petites annonces
Agenda
Boutique
famille chrétienne
CHAQUE MERCREDI, RECEVEZ PAR E-MAIL
LE MEILLEUR DE FAMILLE CHRÉTIENNE
JE M'INSCRIS À LA NEWSLETTER
Nicole et Rég
notre école d
S'ABONNER
Je souhaite poursuivre la lecture de cet article !
Les cookies assurent le bon fonctionnement de nos services. En poursuivant votre navigation vous en acceptez l'utilisation. Pour en savoir plus, rendez-vous ici.
0
100
200
300
400
500

Catalogue

- Introduction
- Dataset
- Model based on HTML content
  - Simple Method
  - Text-CNN
  - Text-RNN
- Model based on screenshots
- Experiments and Results
- Problem and Future Work

Model based on HTML content

- Simple Method
  - Bag of word
    - Binary (Used in paper [4]Web Classification Using Support Vector Machine)
    - Count
    - Tf-Idf
  - Classification Model
    - SVM
    - GaussianNB
    - KNN
    - DecisionTree

Model based on HTML content

- Text-CNN
  - Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

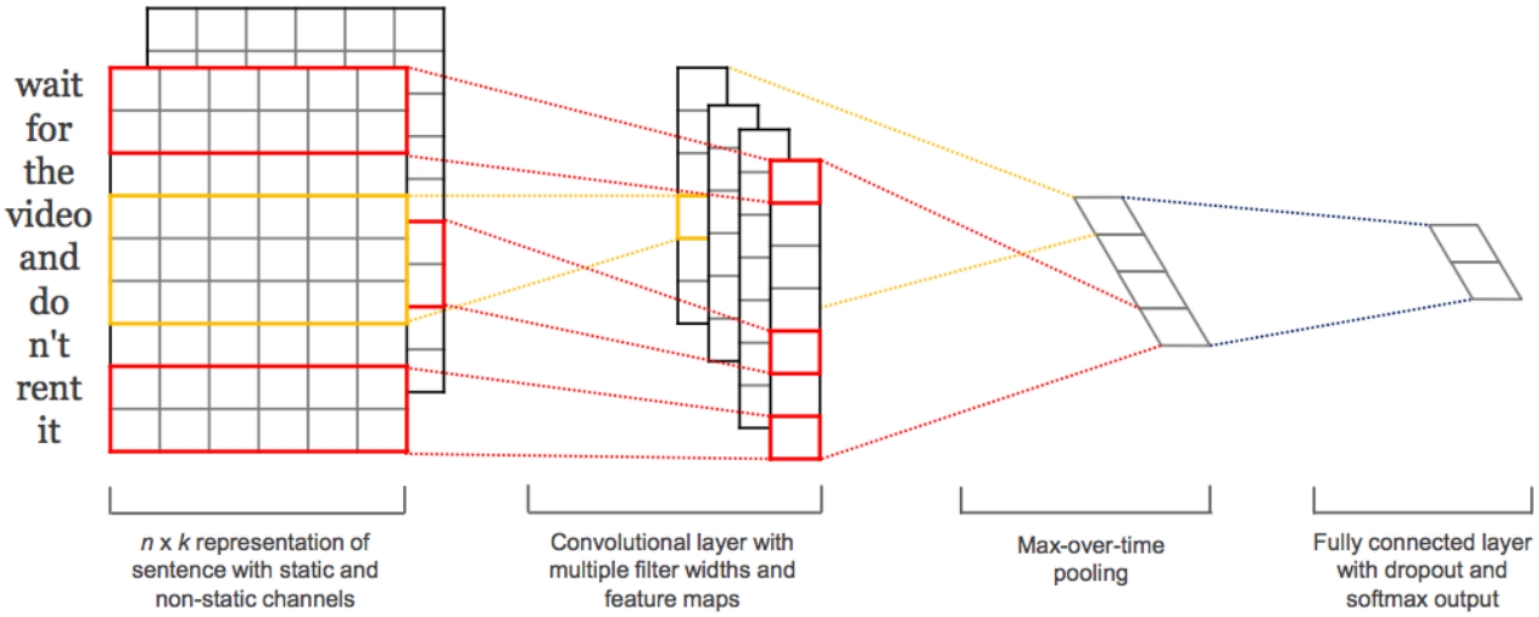


Figure 1: Model architecture with two channels for an example sentence.

Model based on HTML content

- Text-RNN
  - Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 207-212).

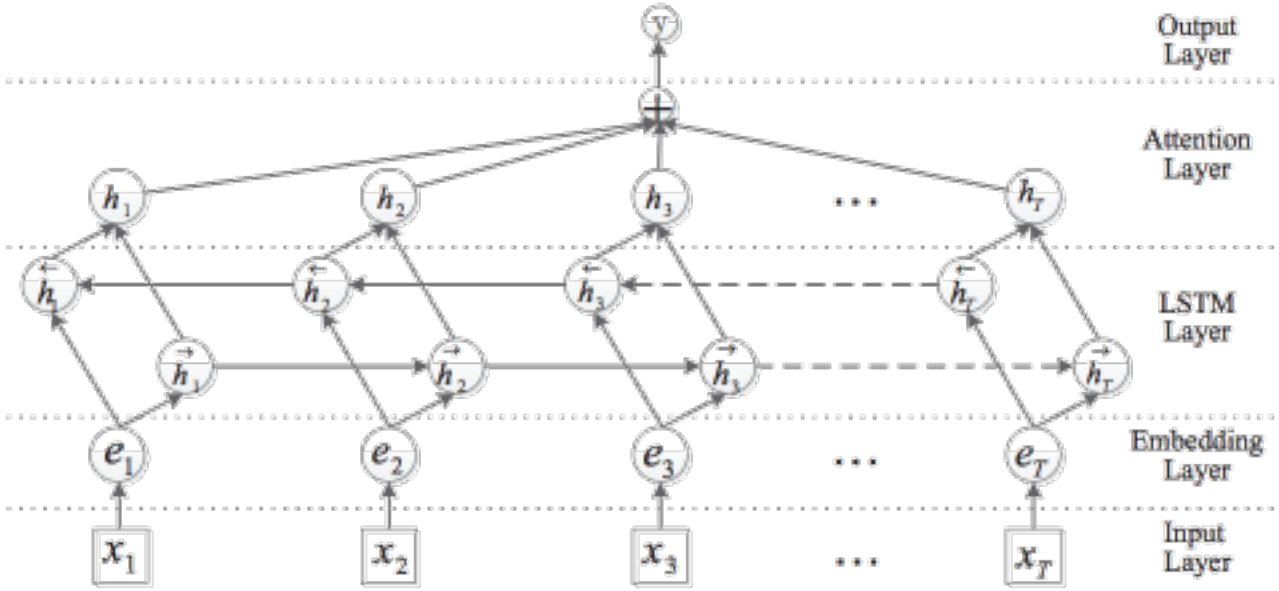


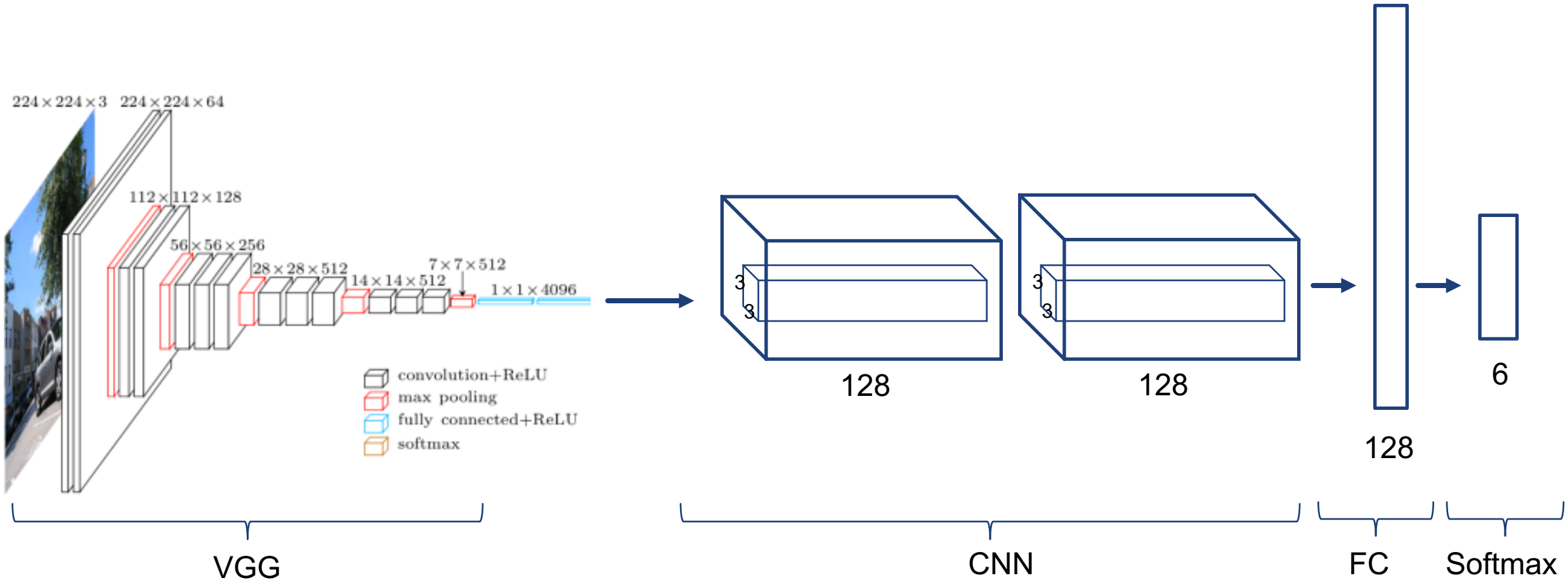
Figure 2: Bidirectional LSTM model with Attention

Catalogue

- Introduction
- Dataset
- Model based on HTML content
- Model based on screenshots
  - VGG+CNN+FC+Softmax
  - Vision based page segmentation
- Experiments and Results
- Problem and Future Work

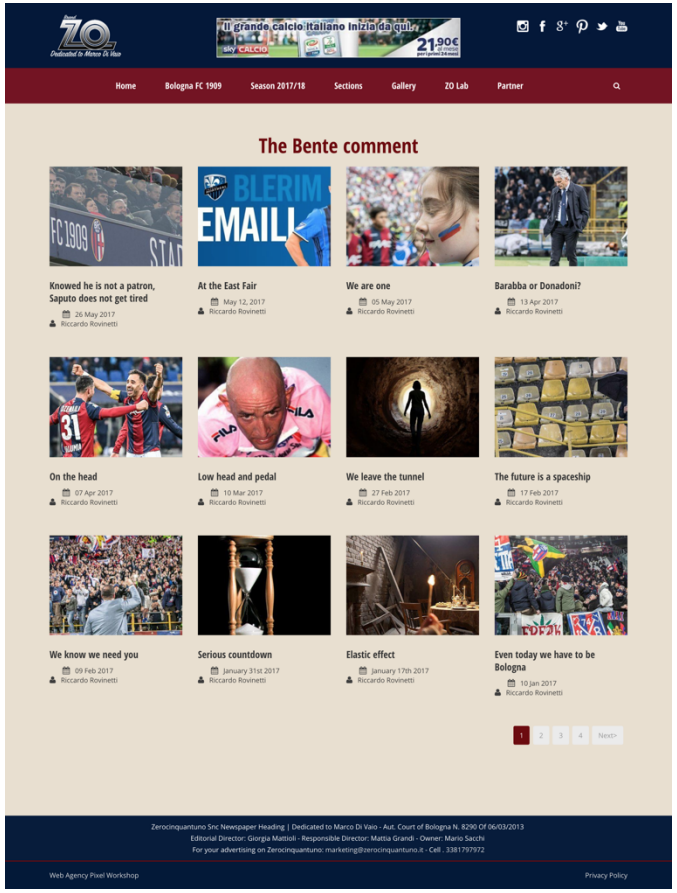
Model based on screenshots

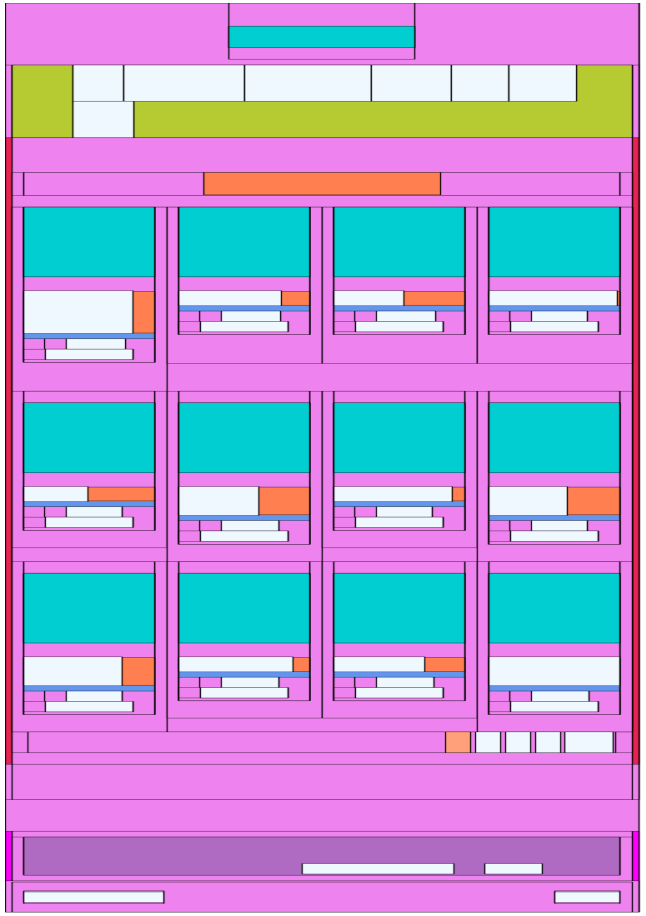
- VGG+CNN+FC+Softmax

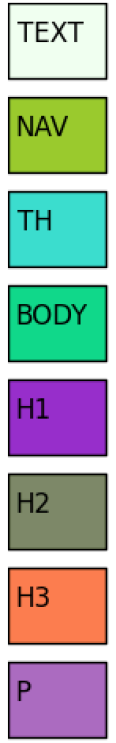


Screenshots to Layout

- Vision based page segmentation
  - Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2003). Vips: a vision-based page segmentation algorithm.







Catalogue

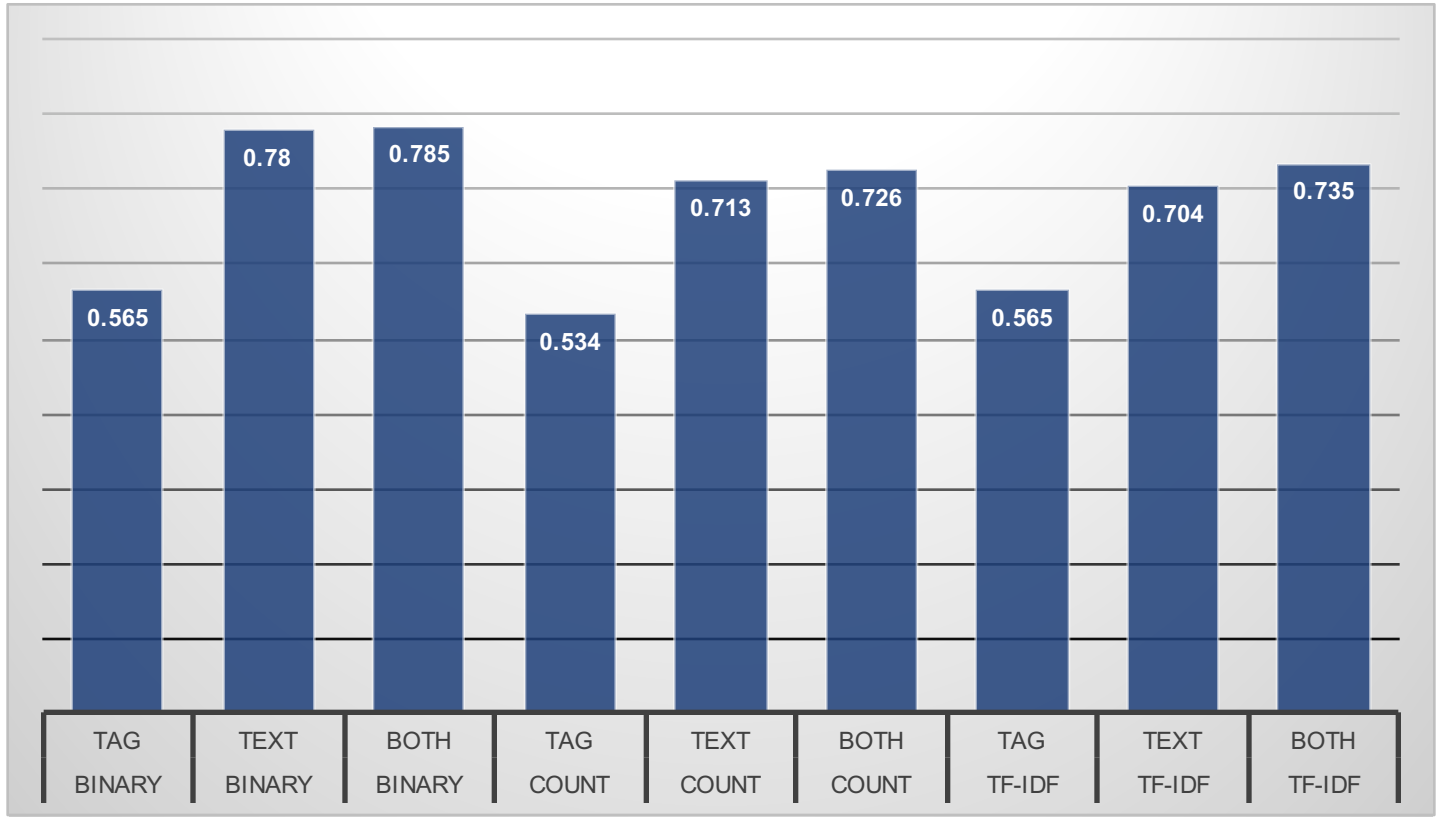
- Introduction
- Dataset
- Model based on HTML content
- Model based on screenshots
- Experiments and Results
  - Simple Method
  - Text-CNN
  - Text-RNN
  - Image-CNN
- Problem and Future Work

Simple Method

- Parameter
- Bag of word: binary/count/tf-idf
- Features:tag/text/both
- Model: SVM/GaussianNB/KNN/DecisionTree

Simple Method

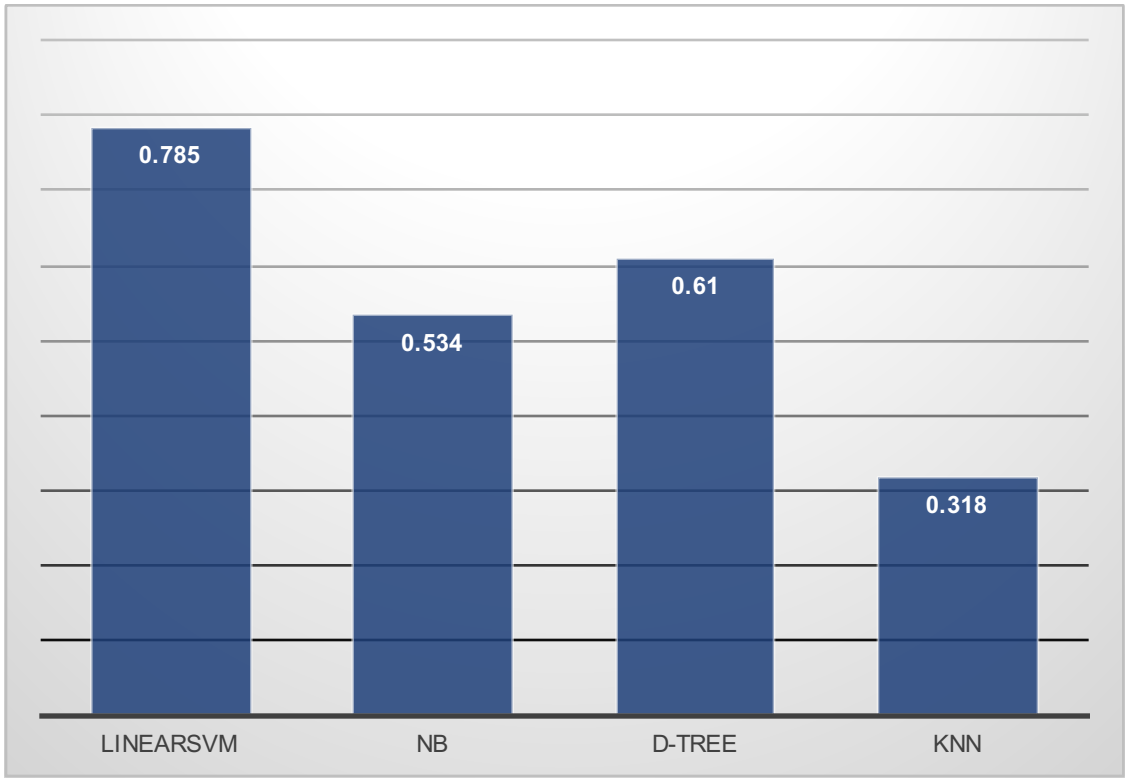
- Different bag of word: binary/count/tf-idf
- Different features:tag/text/both
- Model: Linear-SVM



*The percent of the majority label “Textual” in test set is 42.6%

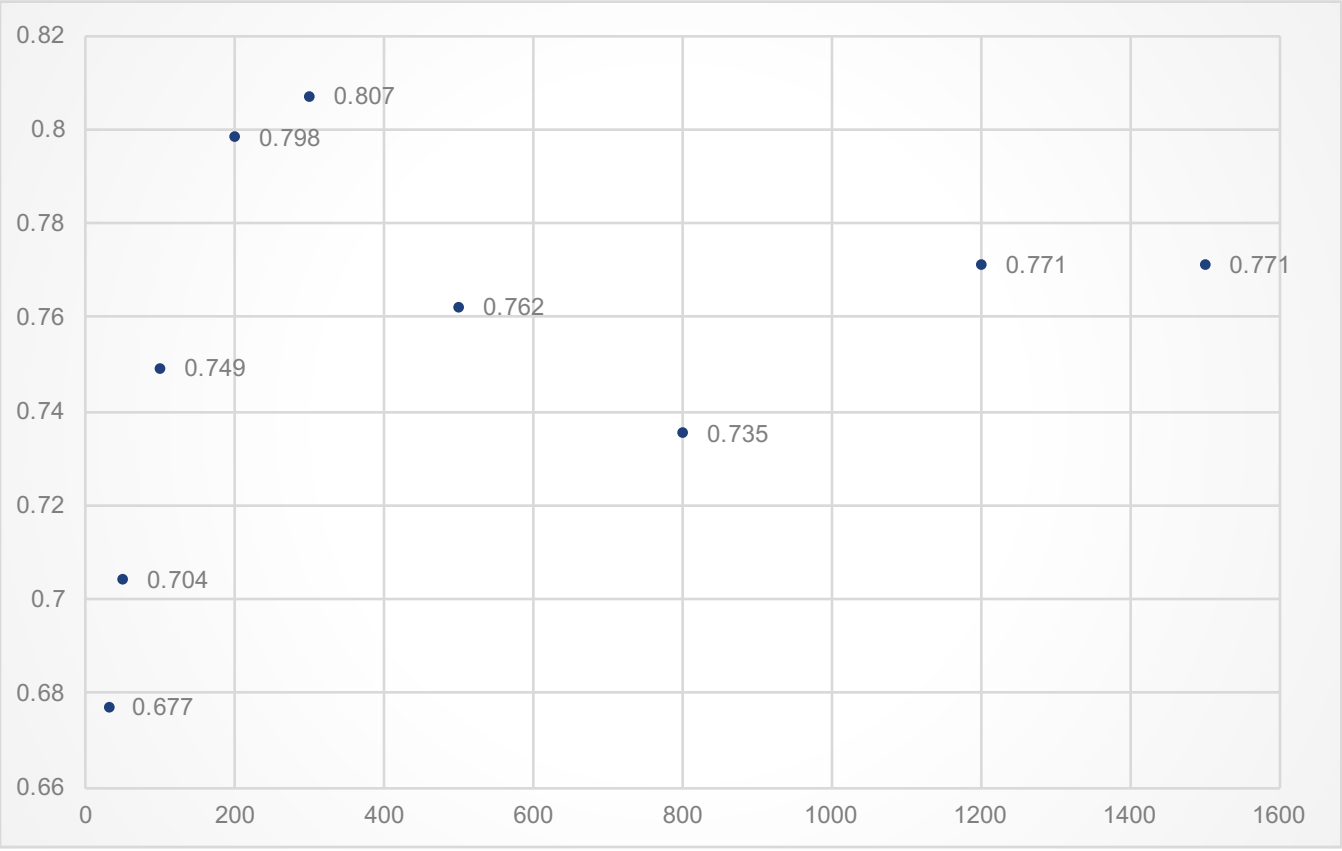
Simple Method

- Different model: SVM/GaussianNB/KNN/DecisionTree
- Bag of word: binary
- Feature: Tag and text



Simple Method+PCA

- Use PCA to reduce dimension based on binary tag and text features
- Accuracy with different feature dimension



Text-CNN

- Parameter
  - Filter size: the size of each filter (For example, in figure 1, filter size is 2,3)
  - Num_filter: the number of fliter for each filter size
  - Max document Length: limit the length of the page (Padding or discard)
  - Feature type: text/tag/tag_with_text

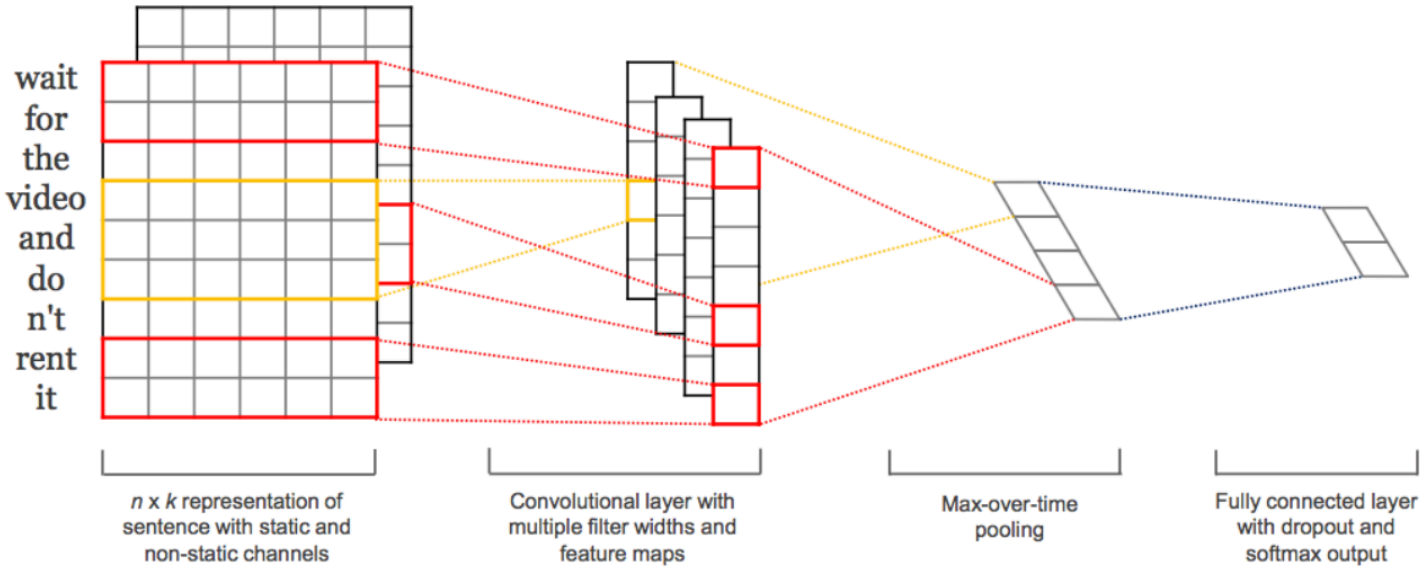
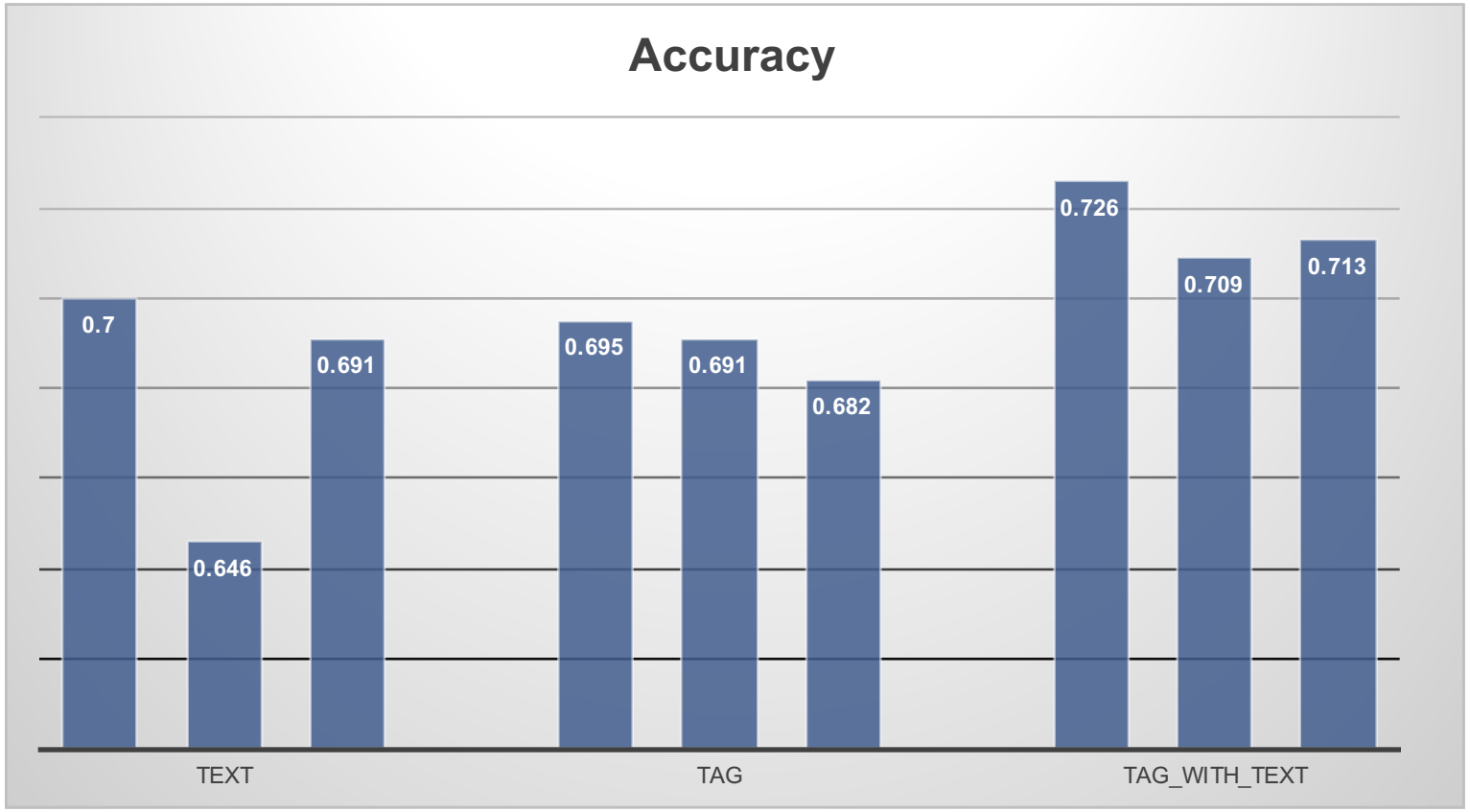


Figure 1: Model architecture with two channels for an example sentence.

Text-CNN

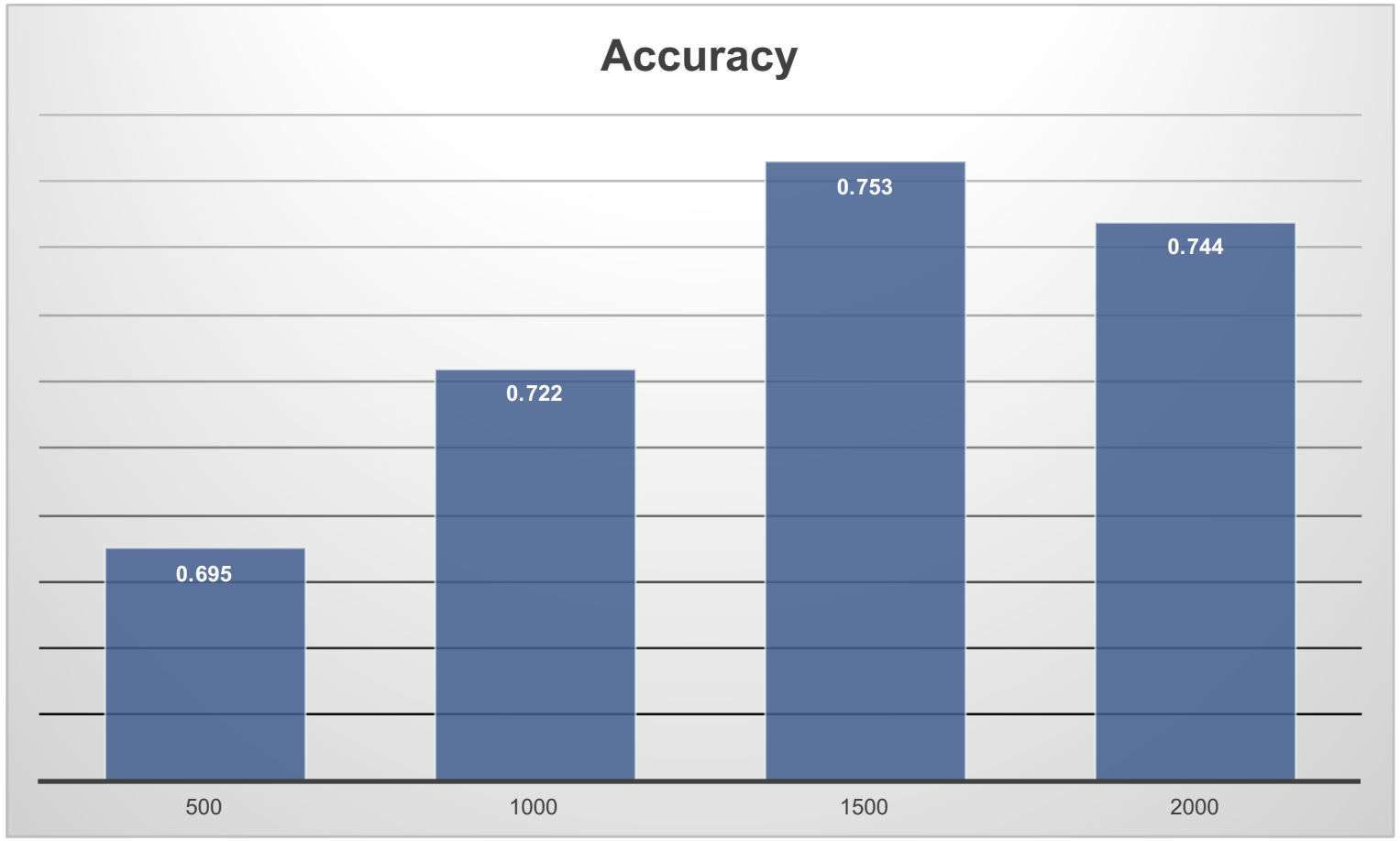
- Different feature type: text/tag/tag_with_text
- Different Fliter size:1,2,3,4,5/3,4,5/1,2,3



Num filter:128
Max document Length:1000

Text-CNN

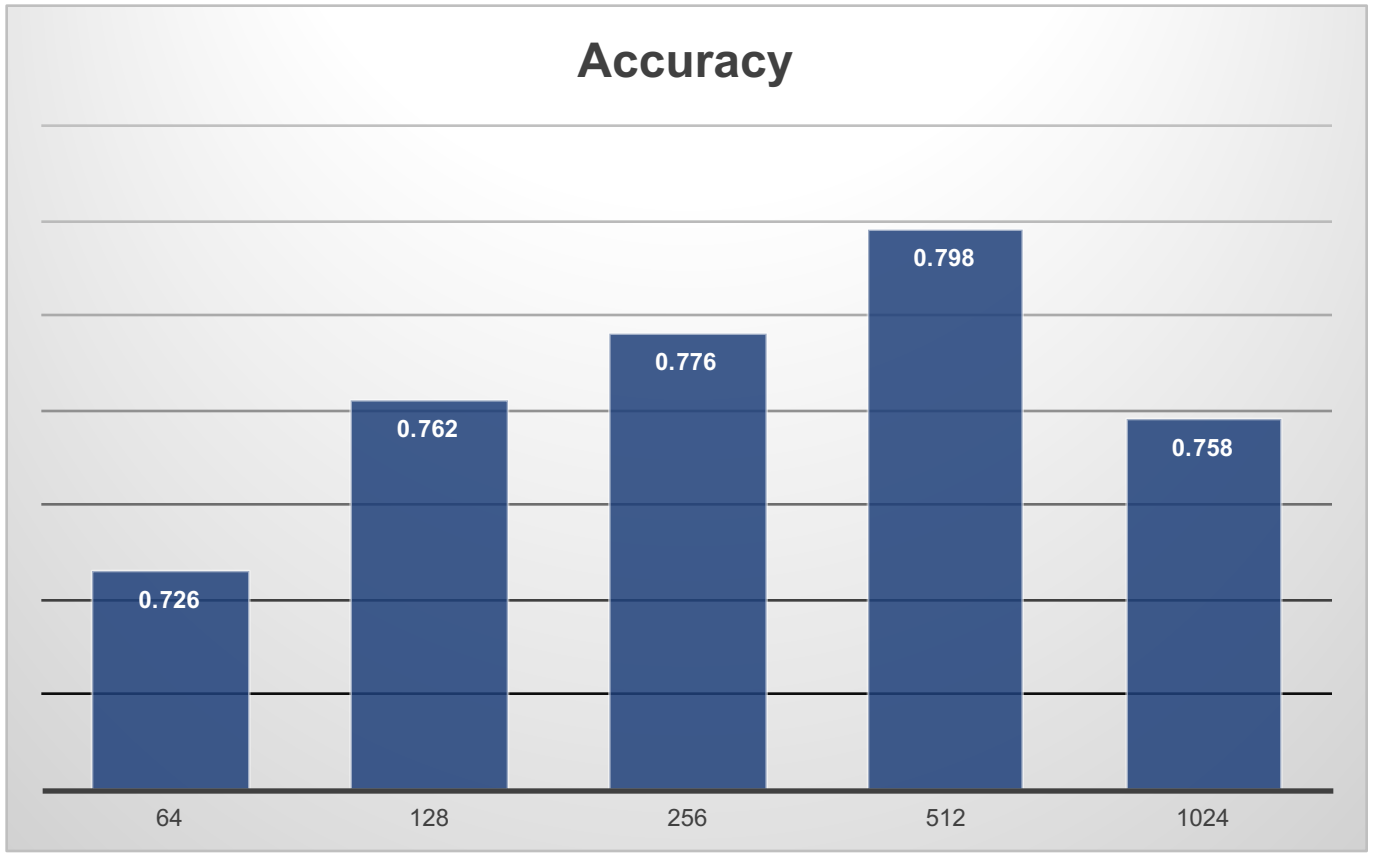
- Different max document length:500/1000/1500/2000
  - The average length of pages “tag with text” is 1781; The max length is 14520.



Feature Type: tag with text
Filter size:1,2,3,4,5
Num filter:128

Text-CNN

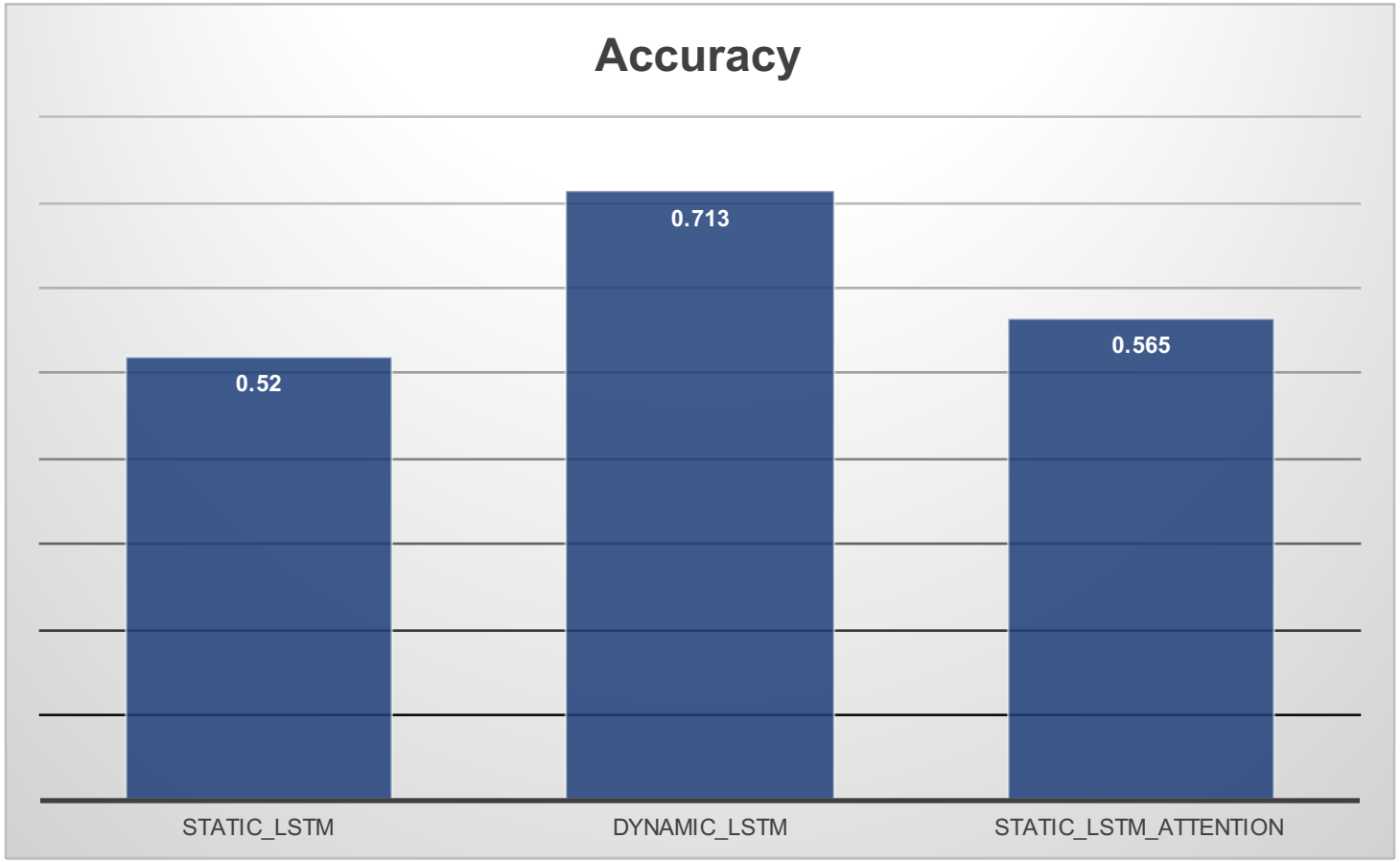
- Different Num fliter: 64/128/256/512/1024



Feature Type: tag with text
Filter size:1,2,3,4,5
Max document Length:1500

Text-RNN

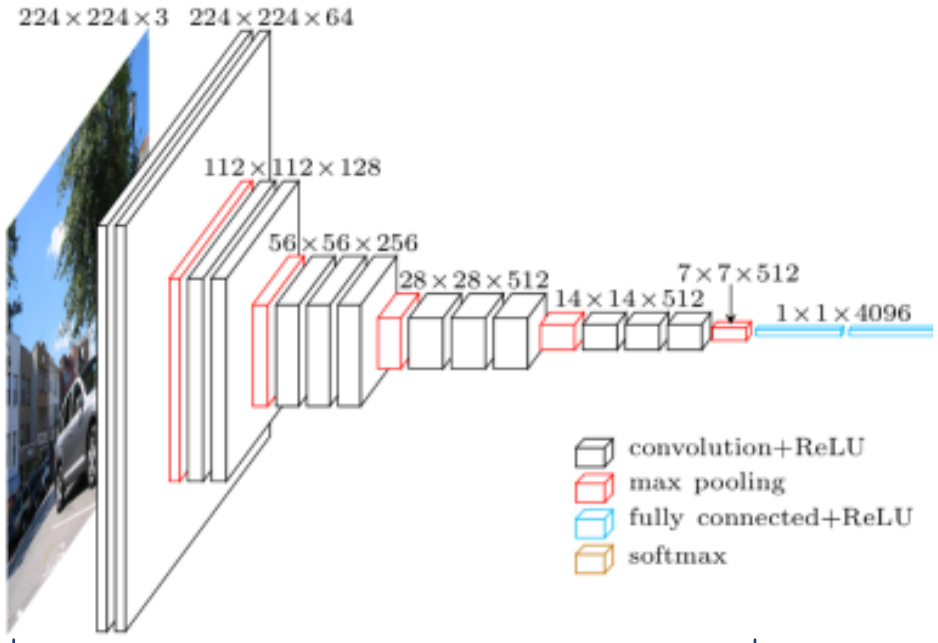
- Different RNN model

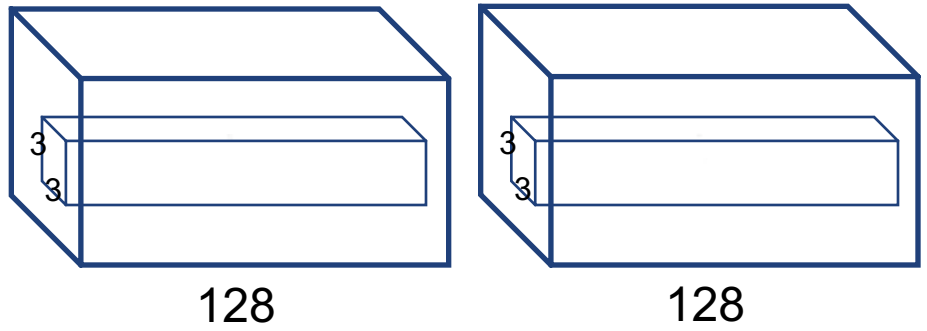


Feature Type: tag with text
Max document Length:1500

Image-CNN

- Parameter:
  - Trainable: Whether vgg layer is trainable or not.
  - Vgg_layer_num: how many vgg layers we keep
  - Cnn_layer_num: how many cnn layer we use after vgg.



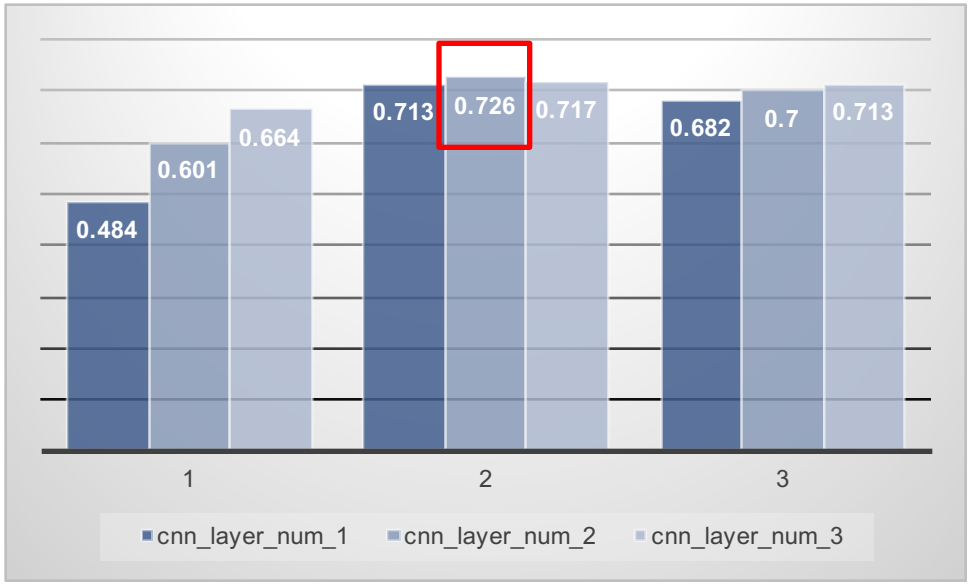


Keep “Vgg_layer_num” layers

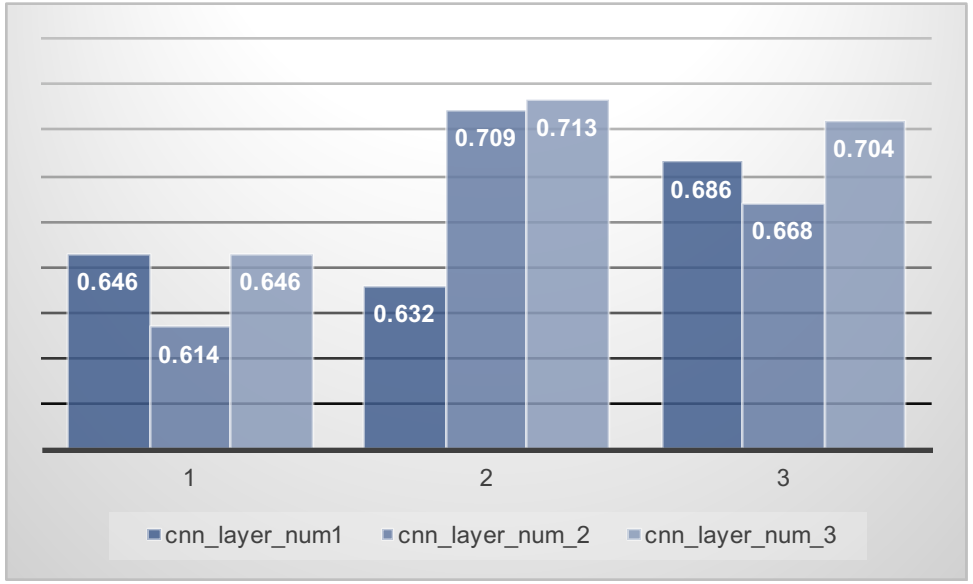
Use “cnn_layer_num” layers

Image-CNN-Screenshot

- Training Time: 800 minutes



Trainable: False



Trainable: True

Compare screenshot and Layout

Screenshot: 2 vgg layers and 2 cnn layers; 800 minutes

average accuracy:0.726

type	precision	recall	F1	TP	FP	TN	FN
Collection	0.658	0.521	0.581	25	13	162	23
Profile	0.614	0.745	0.673	35	22	154	12
Media Item	0.682	0.714	0.698	15	7	195	6
Discussion	1.000	0.333	0.500	1	0	220	2
Other	0.800	0.444	0.571	4	1	213	5
Textual	0.820	0.863	0.841	82	18	110	13

Layout: 2 vgg layers and 2 cnn layers; 800 minutes

average accuracy:0.682

type	precision	recall	F1	TP	FP	TN	FN
Collection	0.900	0.500	0.643	18	2	138	18
Profile	0.643	0.692	0.667	27	15	122	12
Media Item	0.700	0.538	0.609	7	3	160	6
Discussion	0.333	0.333	0.333	1	2	171	2
Other	0.000	0.000	0.000	0	1	168	7
Textual	0.670	0.859	0.753	67	33	65	11

Image-CNN

- Training Time: 2880 minutes (2 days)
- Best Performance: 0.803 with 3 vgg layer and 3 cnn layers, screenshot features.

Summary

Method	Best Performance
LinearSVM	0.785
LinearSVM+PCA	0.807
Text-CNN	0.798
Text-RNN	0.713
Image-CNN	0.803

Catalogue

- Introduction
- Dataset
- Model based on HTML content
- Model based on screenshots
- Experiments and Results
- Problem and Future Work

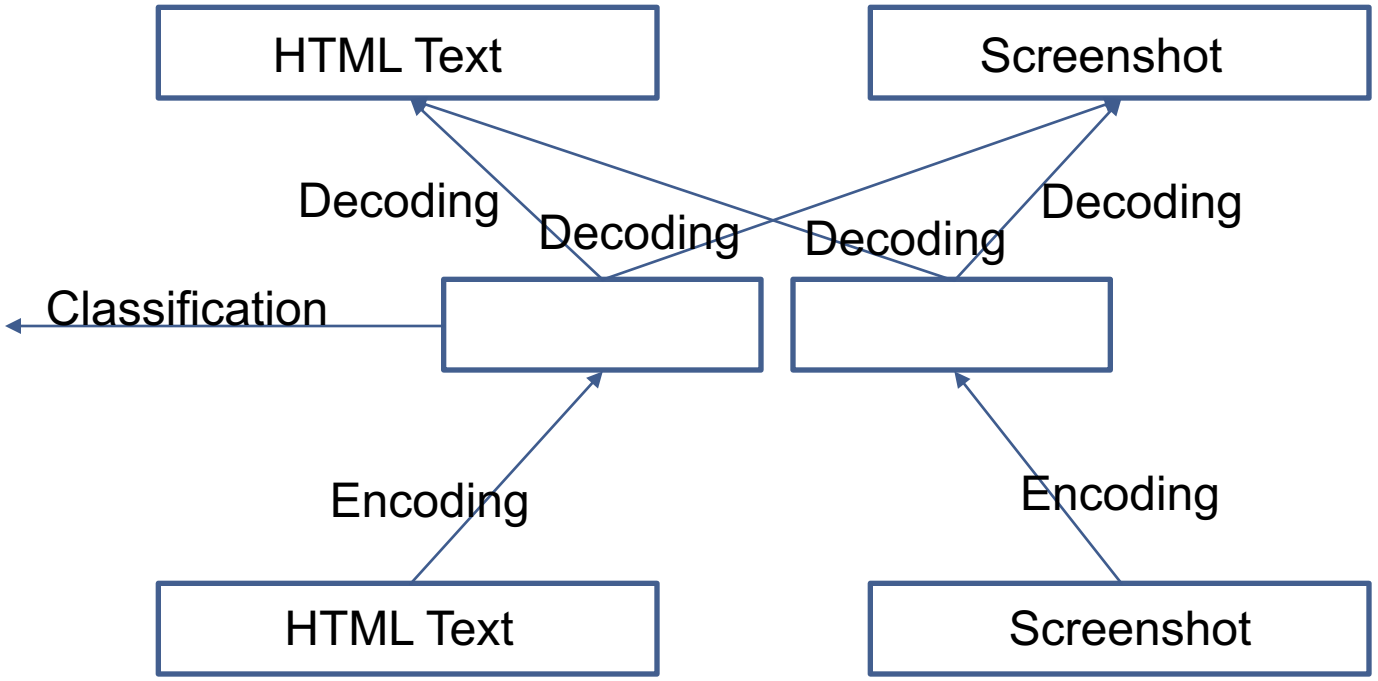
Problem and Future Work

- 1. The performance of Text-RNN is not as good as other method
  - Reason: Long document length and limit training data
  - Possible Solution:
    - Hierarchical Attention Networks
      - Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. J., & Hovy, E. H. (2016). Hierarchical Attention Networks for Document Classification. In *HLT-NAACL* (pp. 1480-1489).
    - Byte pair encoding compression

Method	Best Performance
LinearSVM	0.785
LinearSVM+PCA	0.807
Text-CNN	0.798
Text-RNN	0.713
Image-CNN	0.803

Problem and Future Work

- 2.How to combine text feature and image feature



Problem and Future Work

- 3. Simple method seem to have better performance
  - Reason: Limited labeled data
  - Possible Solution: Unsupervised / Semi-Supervised

Method	Best Performance
LinearSVM	0.785
LinearSVM+PCA	0.807
Text-CNN	0.798
Text-RNN	0.713
Image-CNN	0.803

- Thanks for your Listening
- Q & A