



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

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

DIE KÄFERFARM
Ihr Shop für VW-Käfer-Teile

Suche: Erweiterte Suche »

Home | Warenkorb | Ihr Konto | Neukunde? | Kasse | Katalog Download

Kategorien

Bremse

- Bremssättchen & Beläge
- Bremssättchenbehälter
- Bremsschalter & Gummikappen
- Bremssattel
- Bremsscheiben
- Bremsscheibensets
- Bremstrommeln
- Gummi- und Plastikteile
- Handbremshebel
- Handbremsseile
- Hauptbremszylinder
- Montagezubehör
- Radbremszylinder

Schlüsselelemente und Leitungen

Bremsflüssigkeitsbehälterleitungen (4teilig)
Käfer 12/13/1500 ab 8/67
Artikel-Nr.: 1276
29,95 EUR
(inkl. 19 % MwSt. zzgl. Versandkosten)
Lieferzeit: 4-5 Tage
[In den Warenkorb](#) [Details anzeigen](#)

Bremsflüssigkeitsbehälterleitungen (2teilig)
Käfer 1302/03
Artikel-Nr.: 3353
29,95 EUR
(inkl. 19 % MwSt. zzgl. Versandkosten)
Lieferzeit: 4-5 Tage
[In den Warenkorb](#) [Details anzeigen](#)

Edelstahl Bremschlauch Vorne (Stück)
Käfer bis 7.1964 vorne
Karmann Ghia bis 7.1964 vorne
Bus bis 2.1955 vorne
Artikel-Nr.: 9950
47,20 EUR
(inkl. 19 % MwSt. zzgl. Versandkosten)
Lieferzeit: 4-5 Tage
[In den Warenkorb](#) [Details anzeigen](#)

Edelstahl Bremschlauch vorne (Stück)
Käfer/Karmann Ghia von 8.1964 bis 7.1966 vorne
Bus 3. von 1955 bis 7.1967 vorne
Type 3 Trommelbremse vorne
Artikel-Nr.: 9953

alpha TECHNIK

Presse Impressum Händler Newsletter Kontakt Onlineshop

STARTSEITE PRODUKTE NEWS UNTERNEHMEN RACING GALERIE DOWNLOADS KATALOG ENUMA PRESSE

SIE SIND HIER: > STARTSEITE

Willkommen bei alpha Technik

TRW Moto

OnlineShop ONLINESHOP

Aktuelles

Montag 16. Oktober 2017 alpha Technik Tieferegelung für Yamaha YZF-R3 / MT03 Ab sofort lieferbar für Model... RUBRIK: FAHRWERKSTECHNIK

Montag 16. Oktober 2017 alpha Technik Tieferegelung für Honda CB500 / CBR500 Ab sofort lieferbar für Model... RUBRIK: FAHRWERKSTECHNIK

Montag 02. Oktober 2017 IDM Hockenheim: Reiterteam durch Reiterberger Doppelsieg Das Van-Zen-Rennbahn-BMW Superb... RUBRIK: RENNSPORT

Montag 21. September 2017 alpha Technik Kennzeichenhalter für Yamaha YZF-R3 Ab sofort lieferbar auch für ... RUBRIK: SPECIAL PARTS

Montag 19. September 2017 alpha Technik 35 KW (48 PS) Leistungsänderung für Kymco AK550 Ab sofort lieferbar für Model... RUBRIK: HIGH TECH & QUALITY PARTS

Dienstag 12. September 2017 alpha Technik Tieferegelung für Yamaha YZF-R1 Ab sofort lieferbar für Model... RUBRIK: HIGH TECH & QUALITY PARTS

angebote im Shop

alpha Technik Service

General Games

Categories

- Haunted (11)
- Editors' Picks (52)
- Featured Events (48)
- Shopping Events (47)
- Adventure & Fantasy (44)
- Art (107)
- Bars & Pubs (32)
- Beaches (54)
- Business (7)
- Castles & Ruins (8)
- Chat Hot Spots (16)
- Cosmic (11)
- Cyber (9)
- Discussions & Communities (29)
- Duran Duran (16)
- Education & Nonprofits (59)
- Experiences (10)
- Fashion & Style (323)
- Animations (146)
- Clothing (146)
- Gadgets (21)
- Shoes (15)
- Skins & Shapes (38)
- Hair (21)
- Jewelry (10)
- Tattoos & Accessories (34)
- Games (133)
 - General Games (57)
 - Skill Gaming Regions (62)
 - Help & How To (43)
 - Freebie Spots (4)
 - Creator Resources (14)
 - Newcomer Friendly Spots (20)
 - Sandboxes (8)
 - Home & Garden (99)

Linden Lab uses cookies on our website to, among other things, fulfill user requests (such as enable users to login to our website), provide enhanced functionality for our users (such as user accounts and saved preferences), and enhance web content (so that web content and design are relevant to you and your interests). If you continue to use this site, you agree that we can place these types of cookies on your device. For more information, please review our [Privacy Policy](#).

Like 112 people like this. Sign Up to see what your friends like.

More like General Games: Groups | Events | Shopping

Page 1 of 6

Tyrah and the Curse of the Magical Glythes

Race, fight, outsmart and win with in-world games. Play solo or with friends!

Like 37 people like this. Sign Up to see what your friends like.

Magic Fishing Headquarters

Magic fishing is a new way to earn Linden by fishing. Looking to earn some Linden Dollars? Here you can make friends while earning Linden as you fish and hunt. You need only a free rod to start and one HUD to find locations to fish.

Like 14 people like this. Sign Up to see what your friends like.

Gold Piggy Hunters

Earn free Linden by clicking the gold piggy and teleporting around Second Life to visit new locations. Click the gold piggy and wait five minutes then get paid L\$1. This is a great way for newbies just to earn some free Linden and discover new spots.

Profile

- Structured information about an entity:
 - Products
 - Persons

 **WELDY**
Authorized Partner

Home Produkte Anwendungen Schweisanleitungen & Tipps FAQ's Firma

Search Search DE - EN - FR

Home > Produkte > WELDY Hand-Extruder > WELDY booster EX2



WELDY booster EX2

WELDY booster EX2 - Kunststoff Extrusionsgerät, 230V/3000W

Alle Geräte werden standardmäßig mit EU Stecker ausgeliefert. Falls Sie einen CH-Stecker benötigen, finden Sie den passenden FIX Adapter Schweiz-Europa unter dem generellen Zubehör.

Der professionelle leistungsstarke Hand-Extruder für Kunststoffherstellung und Bauwesen. Der Beste seiner Klasse!

- ▶ Leistung max. 2,2 kg/h Kurz und kompakt
- ▶ Handlich und zuverlässig
- ▶ Lange Lebensdauer
- ▶ Komplett auf Qualität, Funktion und Sicherheit geprüft
- ▶ Problemlos einzufügende Schweißdrähte für flexible Schweißpositionen
- ▶ Arretierbare Taste für Extrusionsschweißen mit wenig Aufwand
- ▶ Produktivität dank schnell austauschbarer Schweißschuhe
- ▶ Drehbarer Griff für bessere Ergonomie
- ▶ Ein auf beiden Seiten montierbarer Knopf sorgt für saubereres Werkzeug

Videos

Bestellnummer: 146.738

Lieferfrist: 5 - 7 Tage

Garantie: 1 Jahr "brine in"



 **Danie/Tschirky**
clean-produkte.ch

SHOP FIRMA DOWNLOADS KONTAKT SERVICE VIDEOS Suchbegriff Warenkorb

Shop > Staub-/Wassersauger / Einscheibenmaschinen >

Nilfisk ALTO Nass-/Trockensauger ATTIX 751 - 11

Der 70 l Innenbehälter, mit dem der Nass-/Trockensauger ATTIX 7 ausgestattet ist, ermöglicht Ihnen lange Arbeitsintervalle ohne Behälterleerung.

Artikelnummer: 302001521
Preis: CHF 1190.00



In den Warenkorb

Haben Sie Fragen zum Produkt? Nehmen Sie unverbindlich [Kontakt](#) mit uns auf.

Detailinformationen **Zubehör**

Top Seller



Nilfisk ALTO Nass-/Trockensauger ATTIX 751 - 11

• Industrielle Spezifikationen, robuste Konstruktion und Leistungsfähigkeit... so definiert man ATTIX 7.
• Das vollautomatische XtremeClean Filterabreinigungssystem (modelabhängig) erleidigt die Filterreinigung regelmäßig, ohne Ihre Arbeit auch nur einen Augenblick zu unterbrechen.
• Waschbarer PET Webefilter für hohe Staubabscheidung und niedrige Wartungskosten.
• Spülbar
• Spitzenleistung, die man sieht, aber nicht hört.
• 70 l Edelstahlbehälter und ein Filtersack für höchste Staubabscheidung und Arbeitseffizienz.
Arbeitsgeräusch (dB(A)) 57



DEVISE: CHF | FRANÇAIS | COMPARER (0) CONTACT MON COMPTE DAILY DEAL BLOG CONNEXION MON PANIER 0

+41 (0)21 566 70 05
Nous répondons du lundi au vendredi de 10h à 18h30

 **Sportmania**
BORN TO RIDE

Online Shop en Suisse pour passionnés de sports de glisse et sports extrêmes!

Rechercher ici... 

KAYAK SKI SKI DE FOND SNOWBOARD LONGBOARD ROLLER TROTINETTE PADDLE - SURF

Accueil > Snowboard K2 Turbo Dream 2017

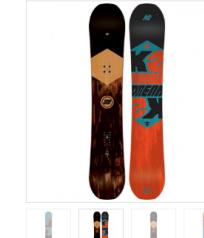
Snowboard K2 Turbo Dream 2017

Ref: K2_turbodream_2017

Une board polyvalente aussi bien pour le freeride que la piste, pour les bons riders qui recherchent de la relance et du pop.

Epuisé 

399,00 CHF



LIVRAISON EN EUROPE
Gratuit pour 200+ d'achats, tous les prix sont TTC, nous ne payez aucun frais de douane, même vers la Suisse

100% SATISFAIT
100% des clients nous ont retourné la marchandise et été remboursés.

DELAI GARANTIS
Les produits en stock sont expédiés sous 24/48h. Nous nous engageons sur la date de livraison indiquée dans votre commande (panier). Délais de livraison par pays



2017   Une question sur ce produit
Ergistrez-vous pour être averti quand le produit sera de nouveau disponible



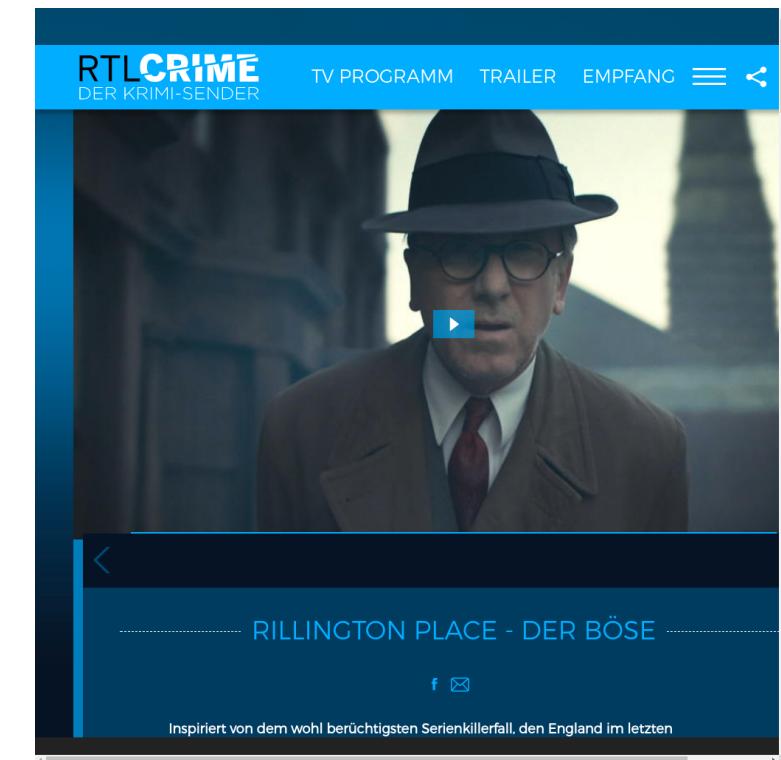
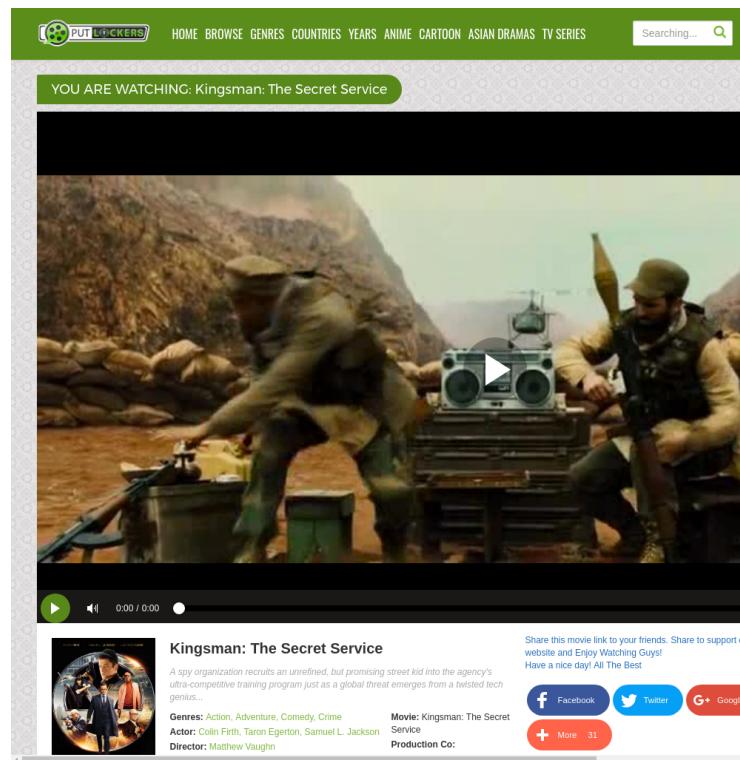


Description Additional Guide des tailles

Snowboard K2 Turbo Dream

Media Item

- Images
- Videos
- Songs



Discussion

- Threads
- Reviews
- Comments

The screenshot shows a forum post on the Banggood.com website. The post is titled "New Free Trial Center Come and try" and has been viewed 8940 times and has 54 replies. It was posted by a user named Yvonne on March 14, 2017, at 22:17. The post includes a large image of a computer monitor displaying a "FREE TRIAL CENTER" advertisement. The ad features a speech bubble saying "Try New And Preview Surprise Of" and a date "20th March". Below the image, there is a message from Yvonne introducing the new trial center and encouraging users to try it. At the bottom of the post, there is a link to a "TRIAL CENTER" page.

Save big on our App! | Affiliate | Sign in or Register My Account

Banggood.com | Forum

Search Posts | Forum Home | Deals | Events | Shopping Guides | General Chat | Order Help | Shop | Blog | Forum Map | Cart

Home / Events

New Topic | Reply | 1 2 ... 6 » | Go!

New Free Trial Center Come and try [copy the Link] [Report]

Views: 8940 | Replies: 54

Post on March 14, 2017 22:17

Yvonne
BG Staff
Replies: 17
Reviews 1
Posts: 4
Send PM

FREE TRIAL CENTER

Try New And Preview Surprise Of

20th March

Hi guys, your favorite Free Trial is coming back.
We have updated the new vision to bring you better service.
Come and try, win super prizes!
If you need help or have any suggestions, please tell us here. We will reply you as soon as possible.

Now let me shortly introduce the new free trial center

TRIAL CENTER - Have a try

SUPPORT

We Are Here To Help

Tell us what you need and your suggestion and we will reply as soon as possible

Textual

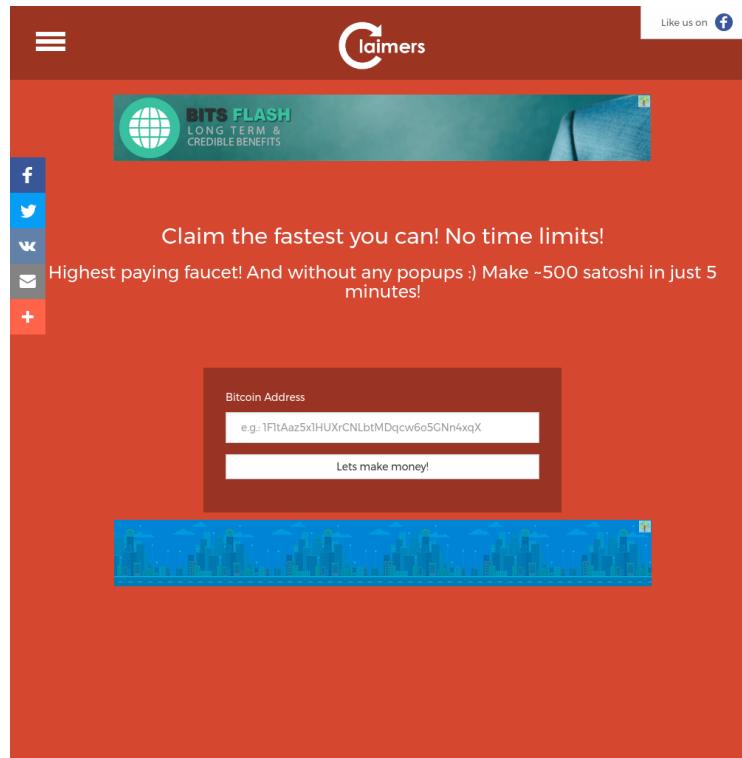
- Newspaper article
- Blog
- Generic information
- Static page

Wie kommt es zu dieser Ablehnung?

Die Theorien darüber, weshalb die Mehrheit der Veterinäre BARF ablehnen scheint, sind vielfältig. In Foren wird heiß spekuliert, ob Tierärzte vielleicht keinen Nutzen von gesunden Tieren hätten und deswegen BARF ablehnen. Schließlich würde Fertigfutter den Hund erst kränklich machen und somit den Weg für eine dauerhafte Behandlung ebnen. Oder ob sie statt von der Futtermittellobby geschmierte Marionetten seien, die sich an Spezialdiäten eine goldene Nase verdienten. Auch diese Liste ist übrigens endlos und meiner Meinung nach ebenso falsch. Ich persönlich bin davon überzeugt, dass die meisten Abiturienten,

Other

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



Flüge > Fluggesellschaften > Emirates

Emirates Flüge günstig buchen mit Opodo

Flüge Hotels Flug + Hotel Autos

Abreiseort: [] Nach: []

Abflug: [] Rückflug: []

1 Erwachsener

Flug suchen Flug + Hotel suchen

Beliebte Emirates Flugrouten

Frankfurt - Bangkok	Flug ab 490€	Prag - Bangkok	Flug ab 595€
Zürich - Dubai	Flug ab 497€	London - Kapstadt	Flug ab 616€
Beirut - Dubai	Flug ab 187€	Auckland - Melbourne	Flug ab 254€

Weitere Ergebnisse

Emirates

Übersetzung aus dem Thailändischen ins Deutsche Sprache

THIS IS NOT A TYPICAL DATING SITE Just send a message and start dating [VIEW PROFILES](#)

Thai virtuelle Tastatur Online

Übersetzen oder menschliche Übersetzung

Deutsch Sauber

Übersetzen mit Hilfe von Google Bing Glosbe Bab.la

THIS IS NOT A TYPICAL DATING SITE Just send a message and start dating • NO FAKE PROFILES • NO CREDIT CARD REQUIRED • GIRL WILL MAKE FIRST MOVE [VIEW PROFILES](#)

Romance Tale

Facebook Twitter Drucken Google+ Mehr...

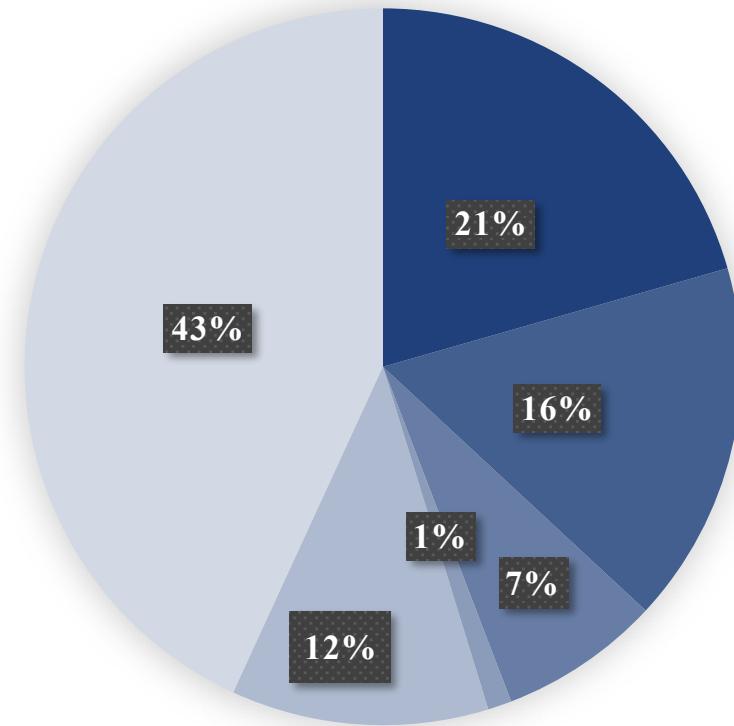
Catalogue

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

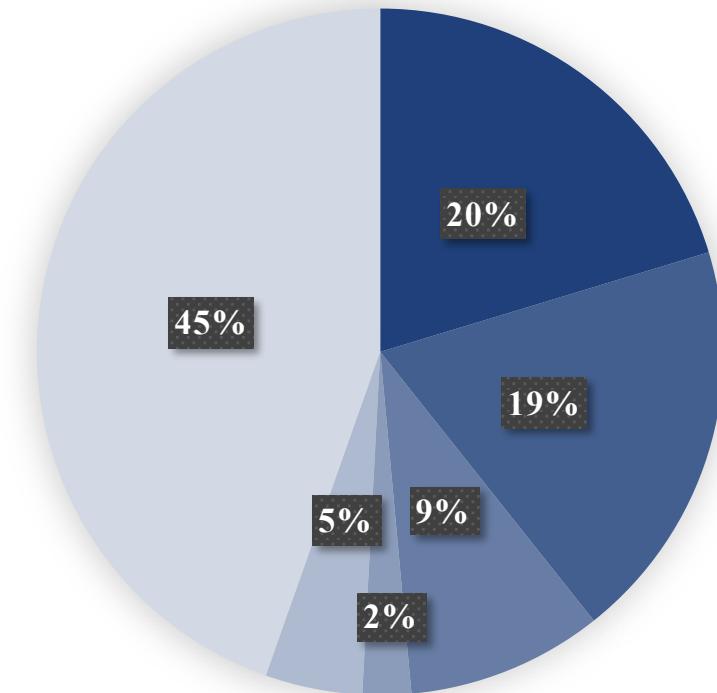
Dataset - Totally

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

Dataset - Labeled

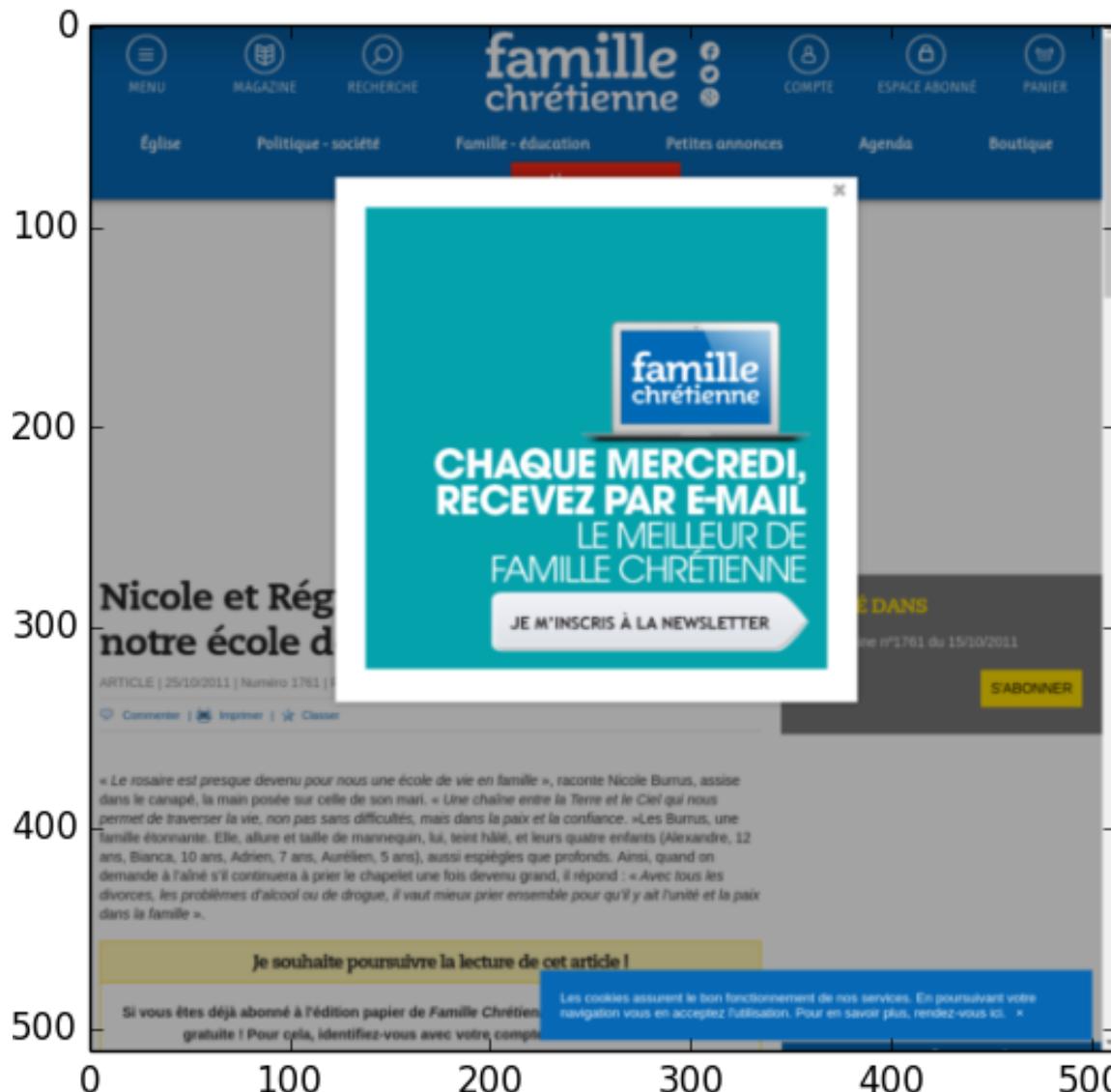


Labeled dataset – Totally
Size: 2791



Labeled dataset – Clean
Size: 1137

- Collection
- Profile
- Media Item
- Discussion
- Other
- Textual



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*.

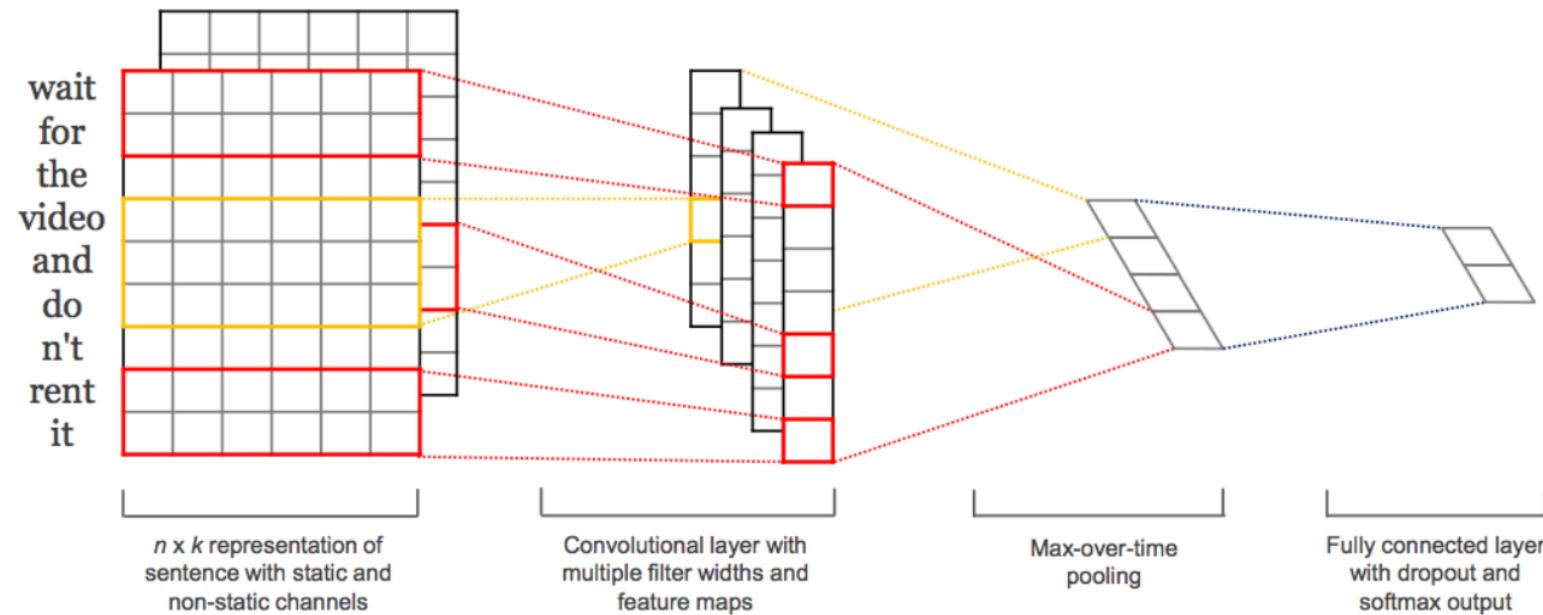


Figure 1: Model architecture with two channels for an example sentence.

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).

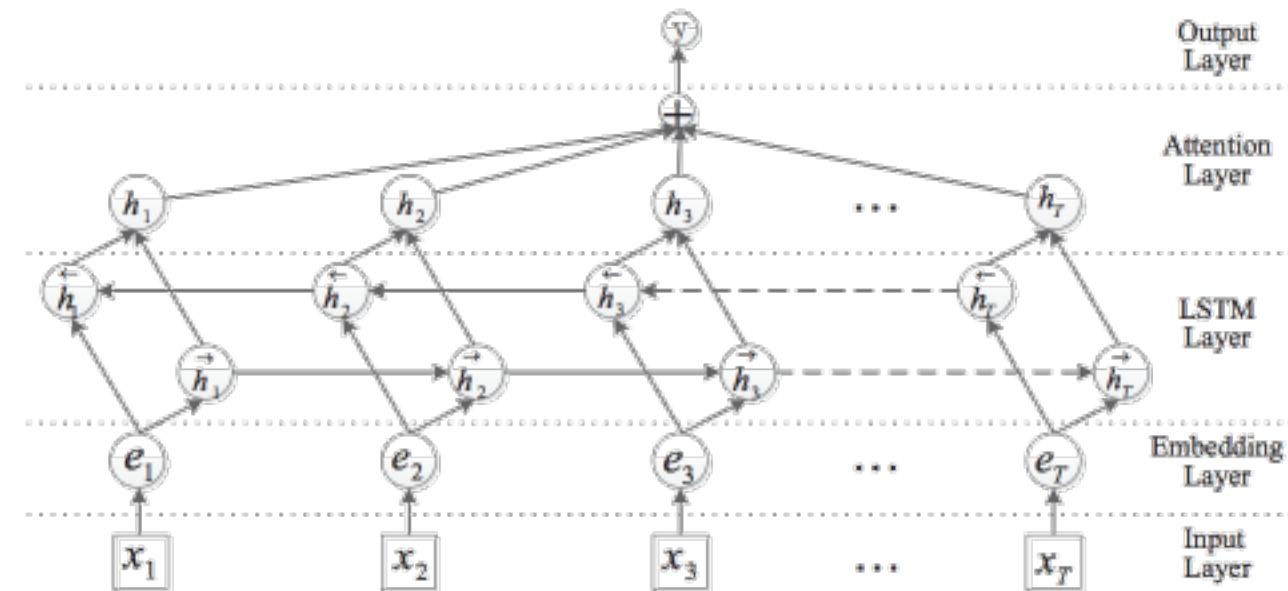


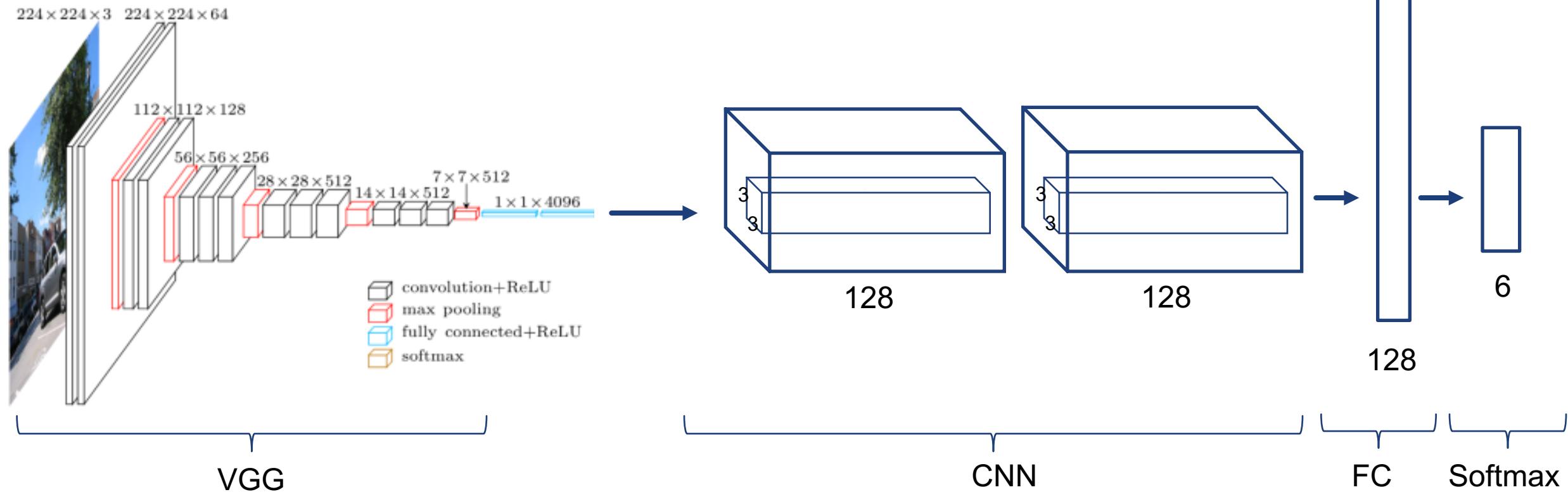
Figure 2: Bidirectional LSTM model with Attention

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

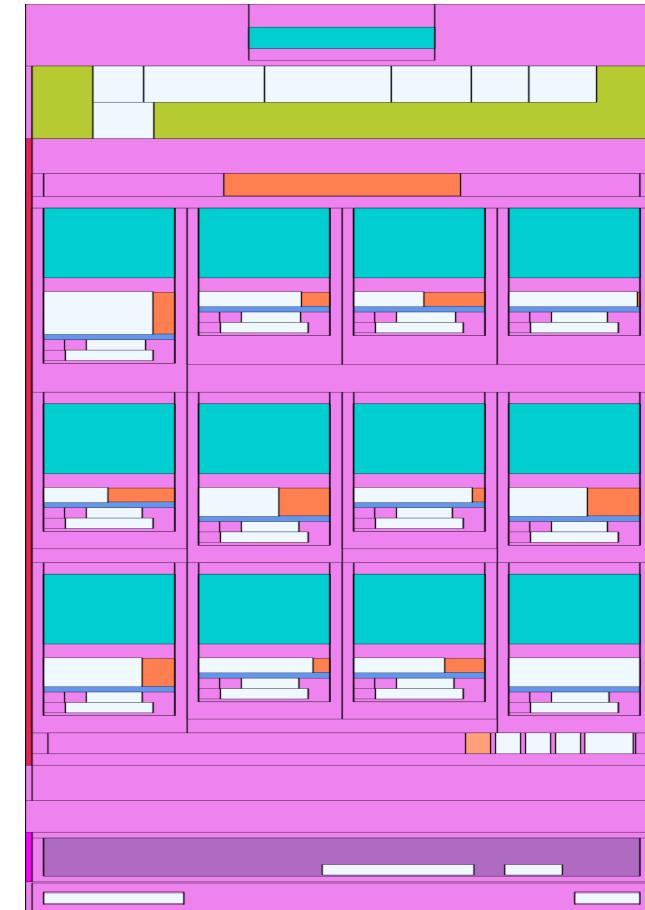
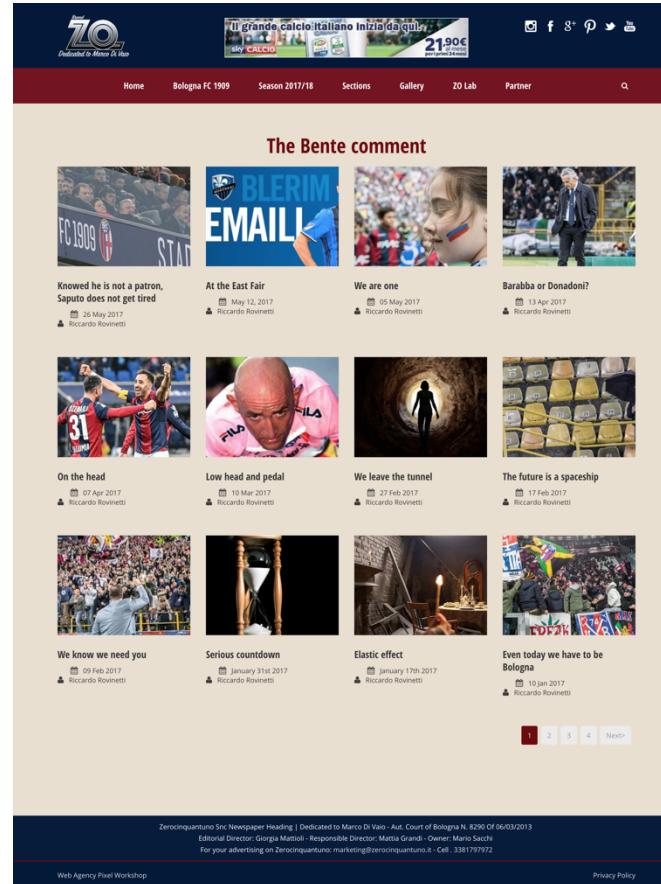
Model based on screenshots

- 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

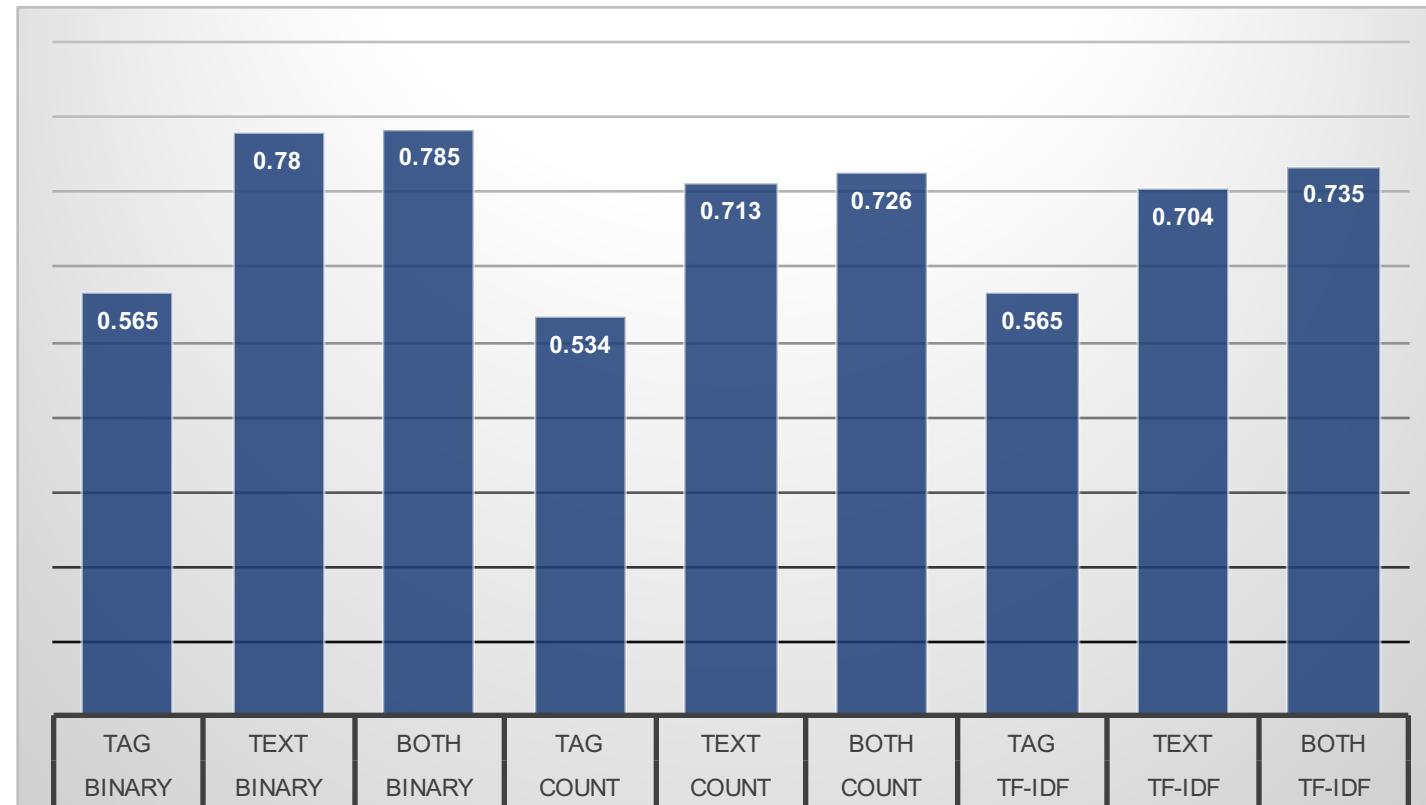
- Introduction
- Dataset
- Model based on HTML content
- Model based on screenshots
- Experiments and Results
 - Simple Method
 - Text-CNN
 - Text-RNN
 - Image-CNN
- Problem and Future Work

Simple Method

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

Simple Method

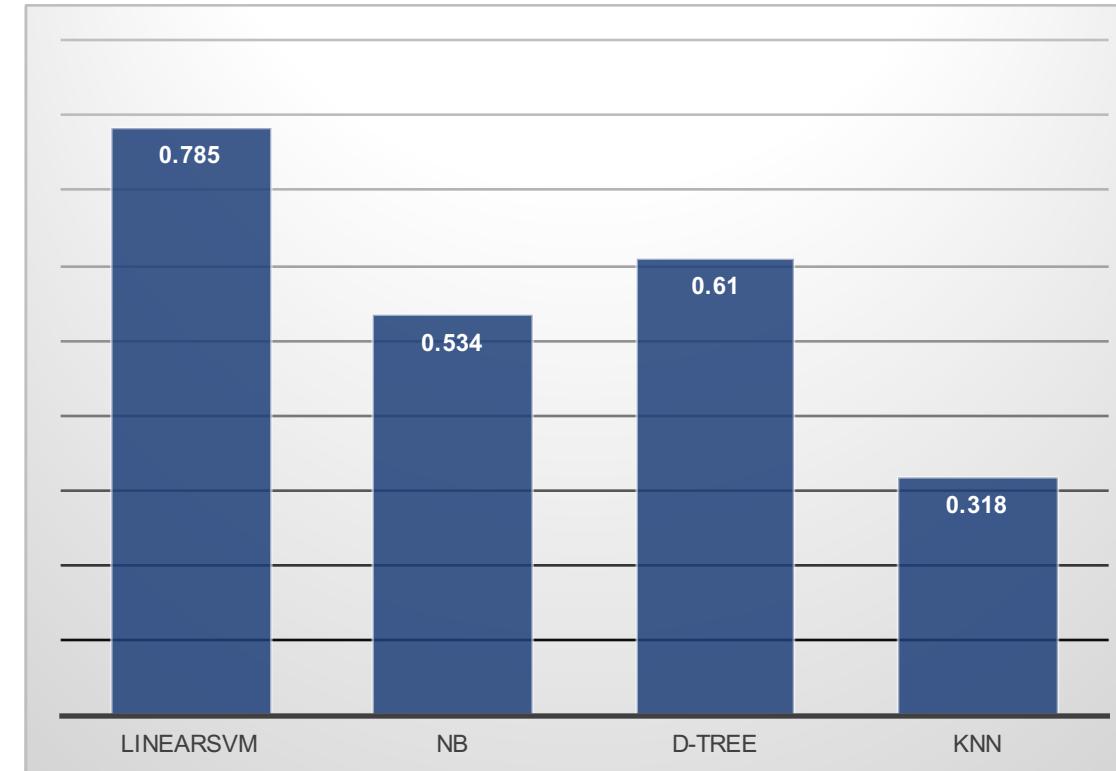
- Different bag of word: binary/count/tf-idf
- Different features: tag/text/both
- Model: Linear-SVM



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

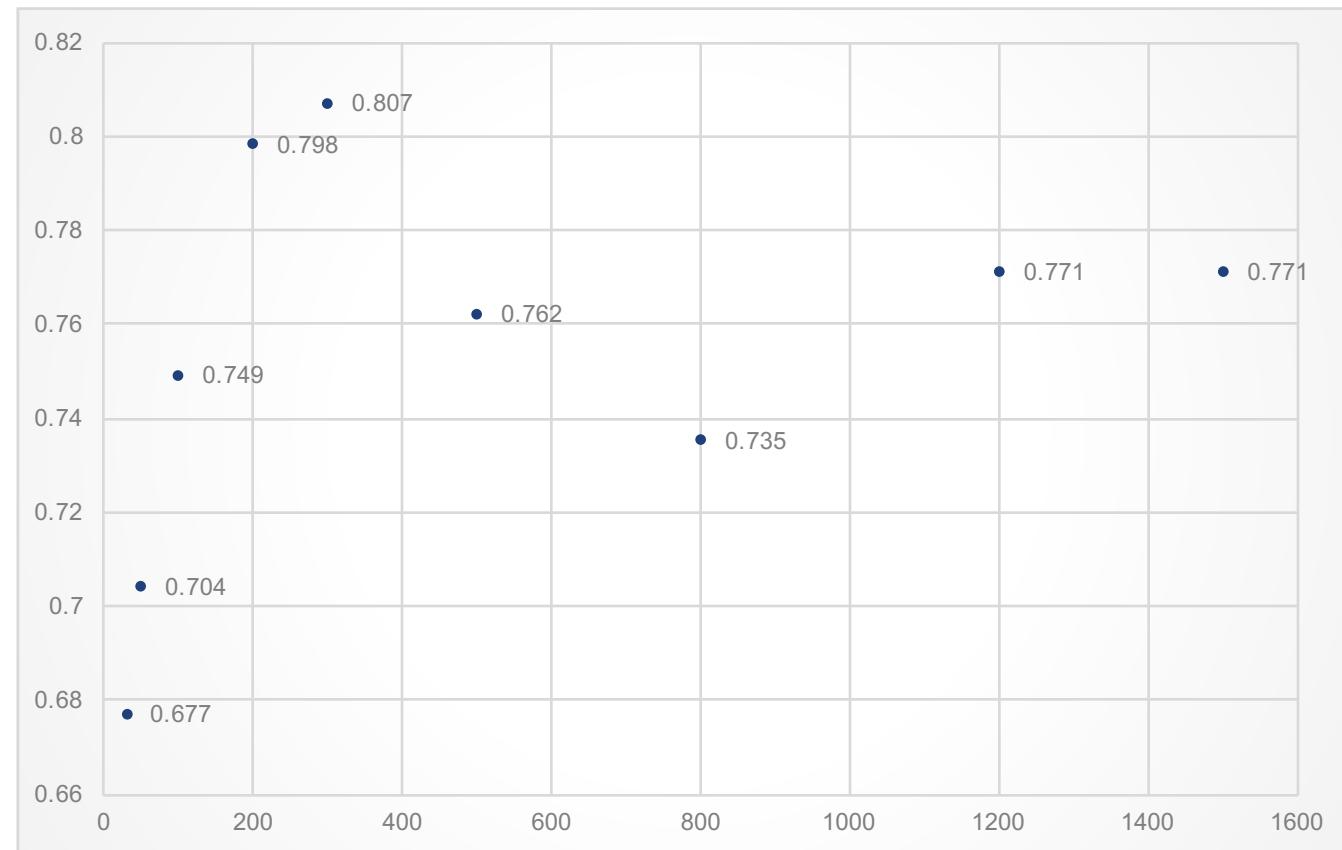
Simple Method

- 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 filter for each filter size
- Max document Length: limit the length of the page (Padding or discard)
- Feature type: text/tag/tag_with_text

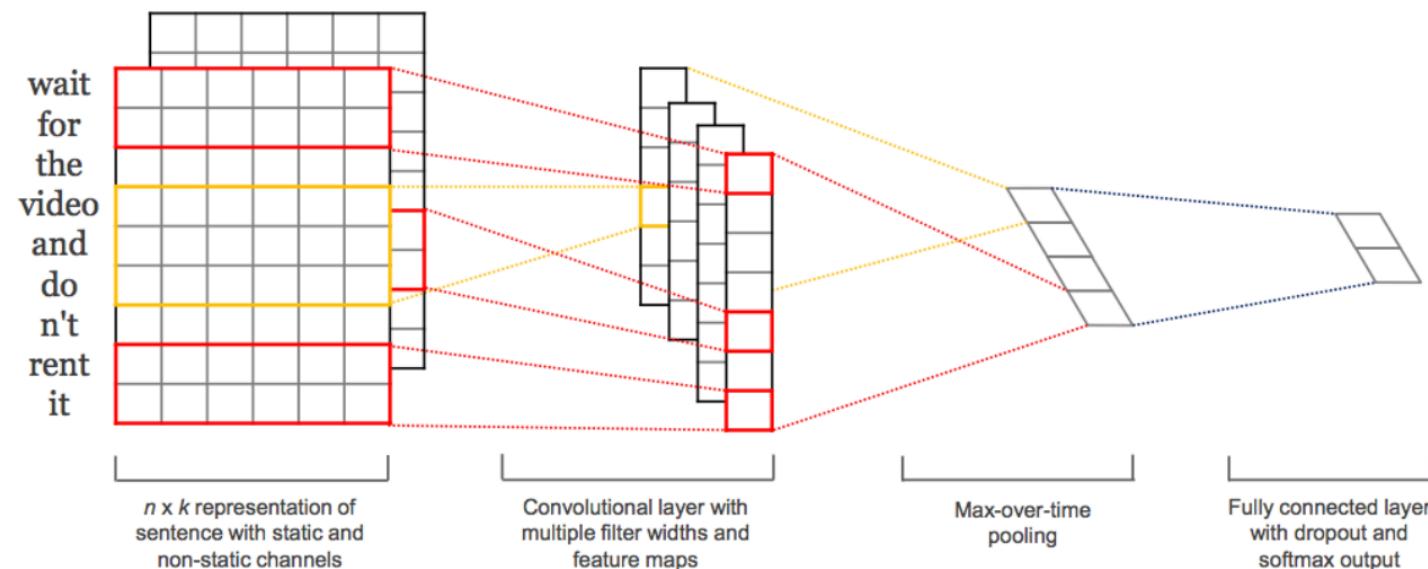
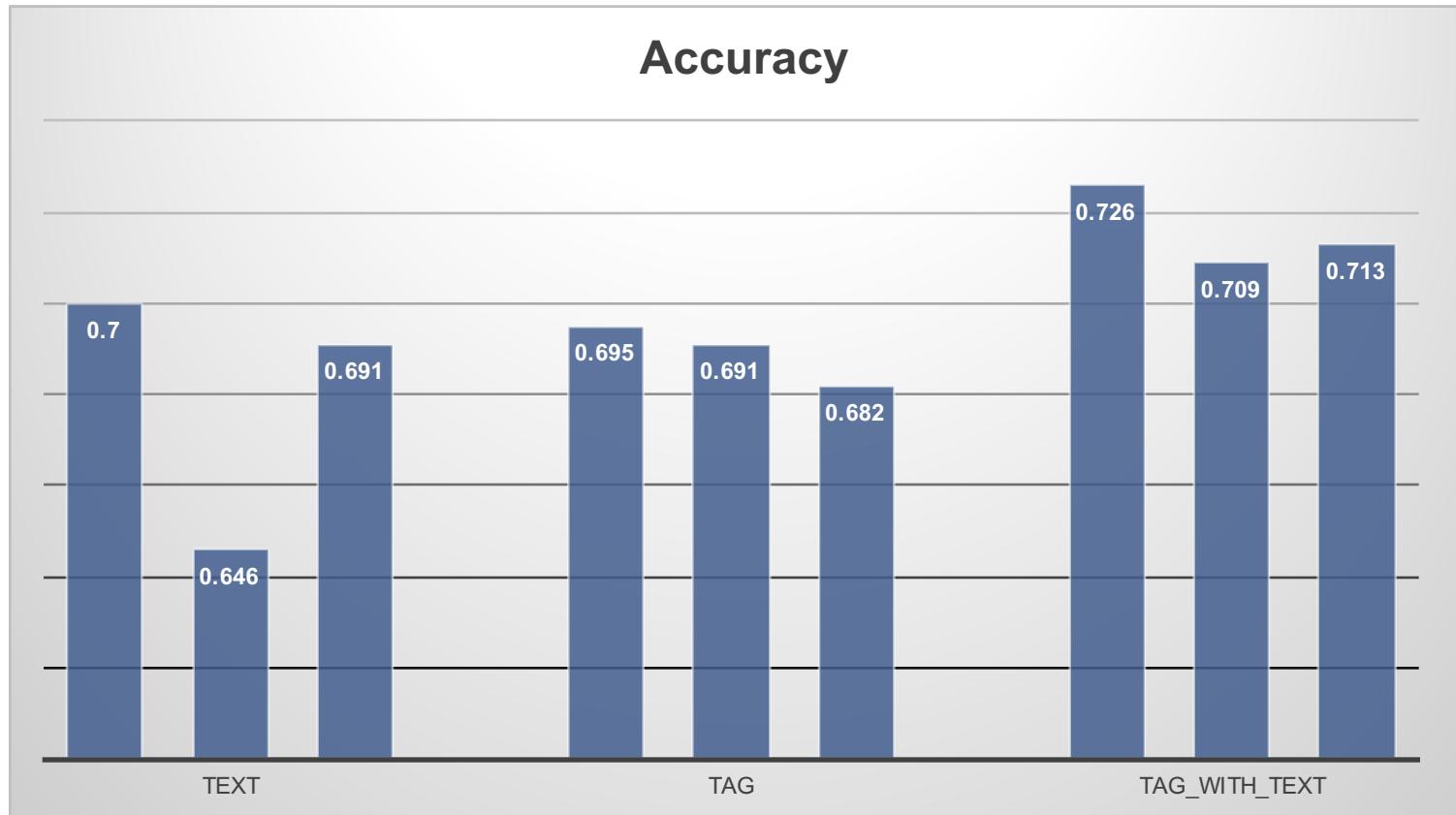


Figure 1: Model architecture with two channels for an example sentence.

Text-CNN

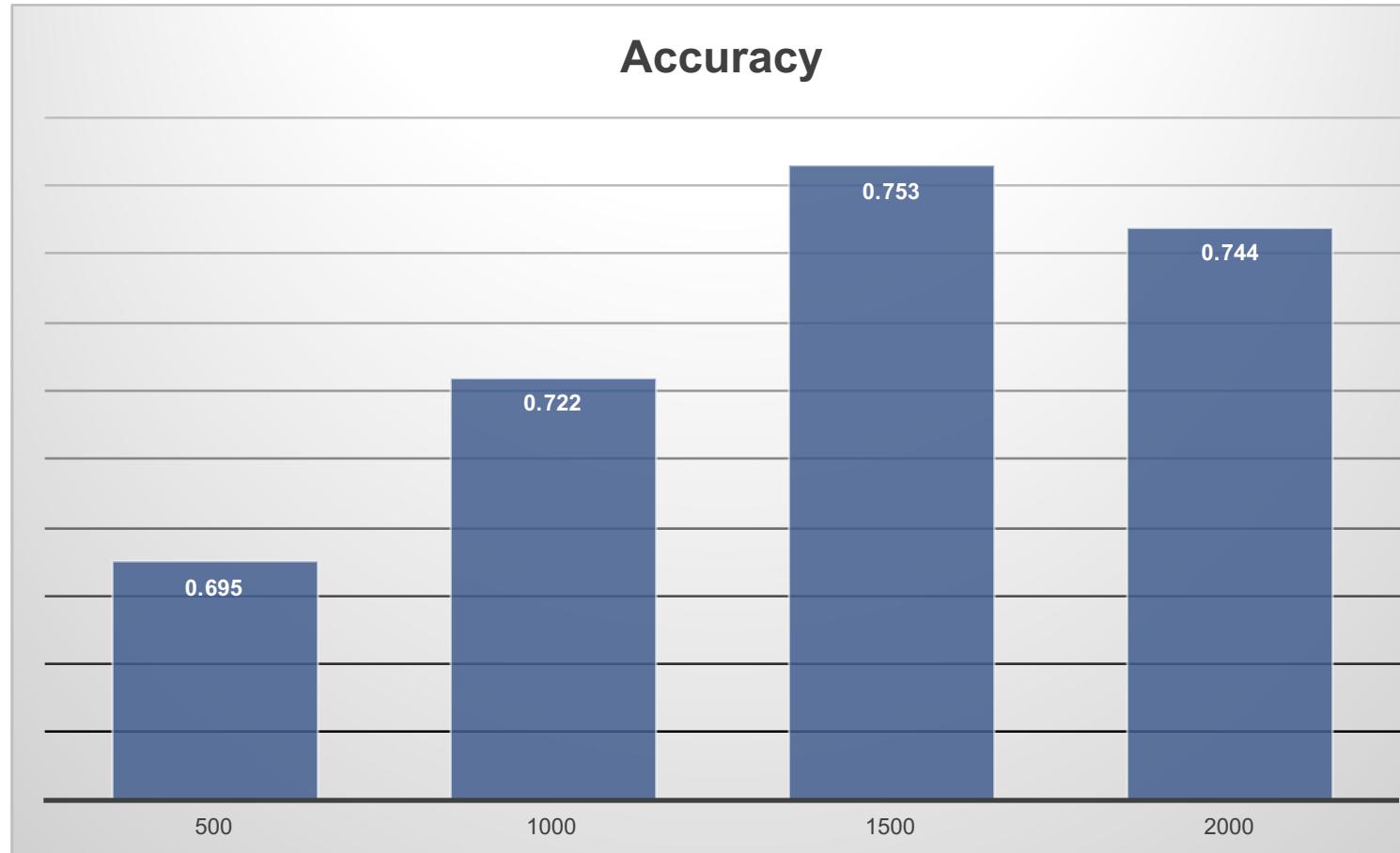
- Different feature type: text/tag/tag_with_text
- Different Filter size: 1,2,3,4,5/3,4,5/1,2,3



Num filter: 128
Max document Length: 1000

Text-CNN

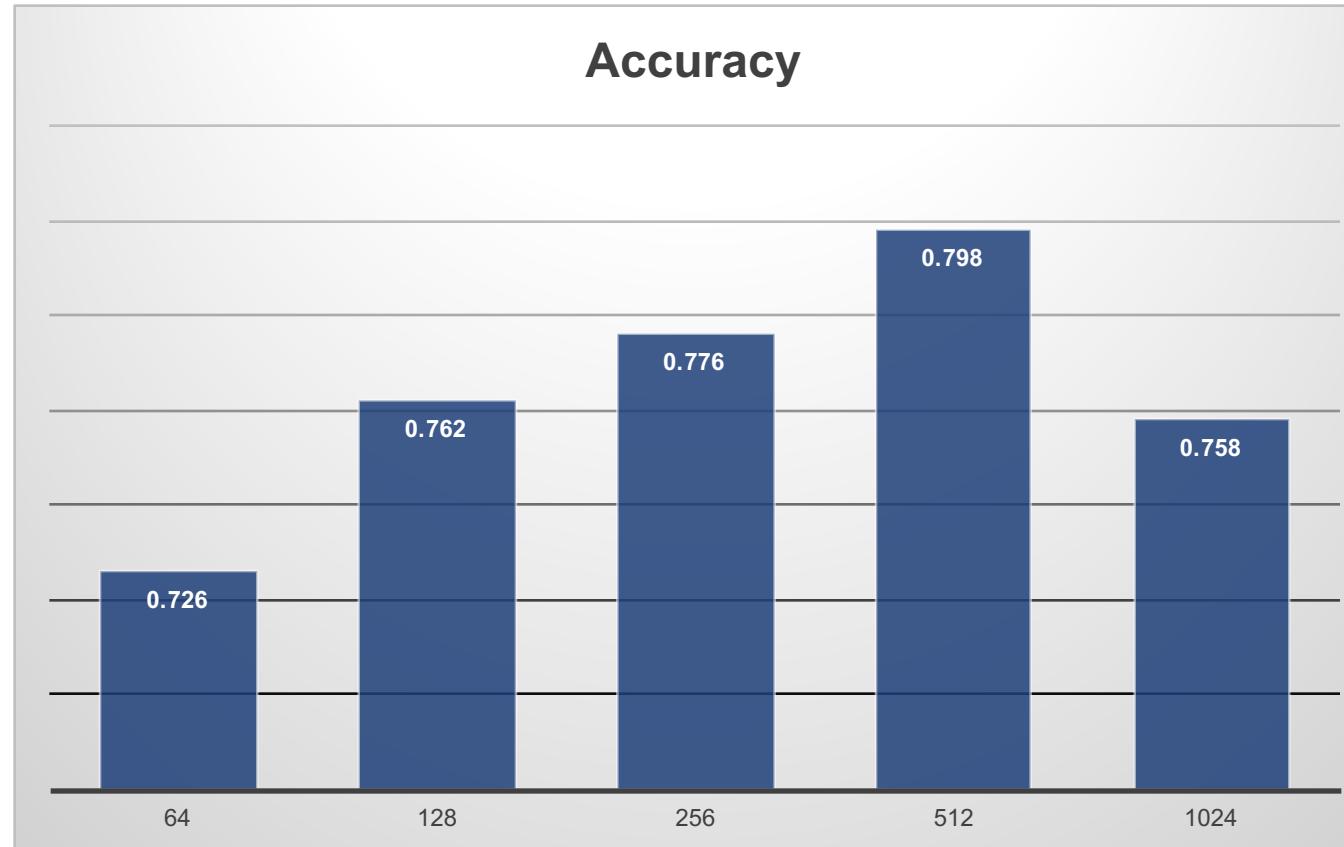
- Different max document length:500/1000/1500/2000
 - The average length of pages “tag with text” is 1781; The max length is 14520.



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

Text-CNN

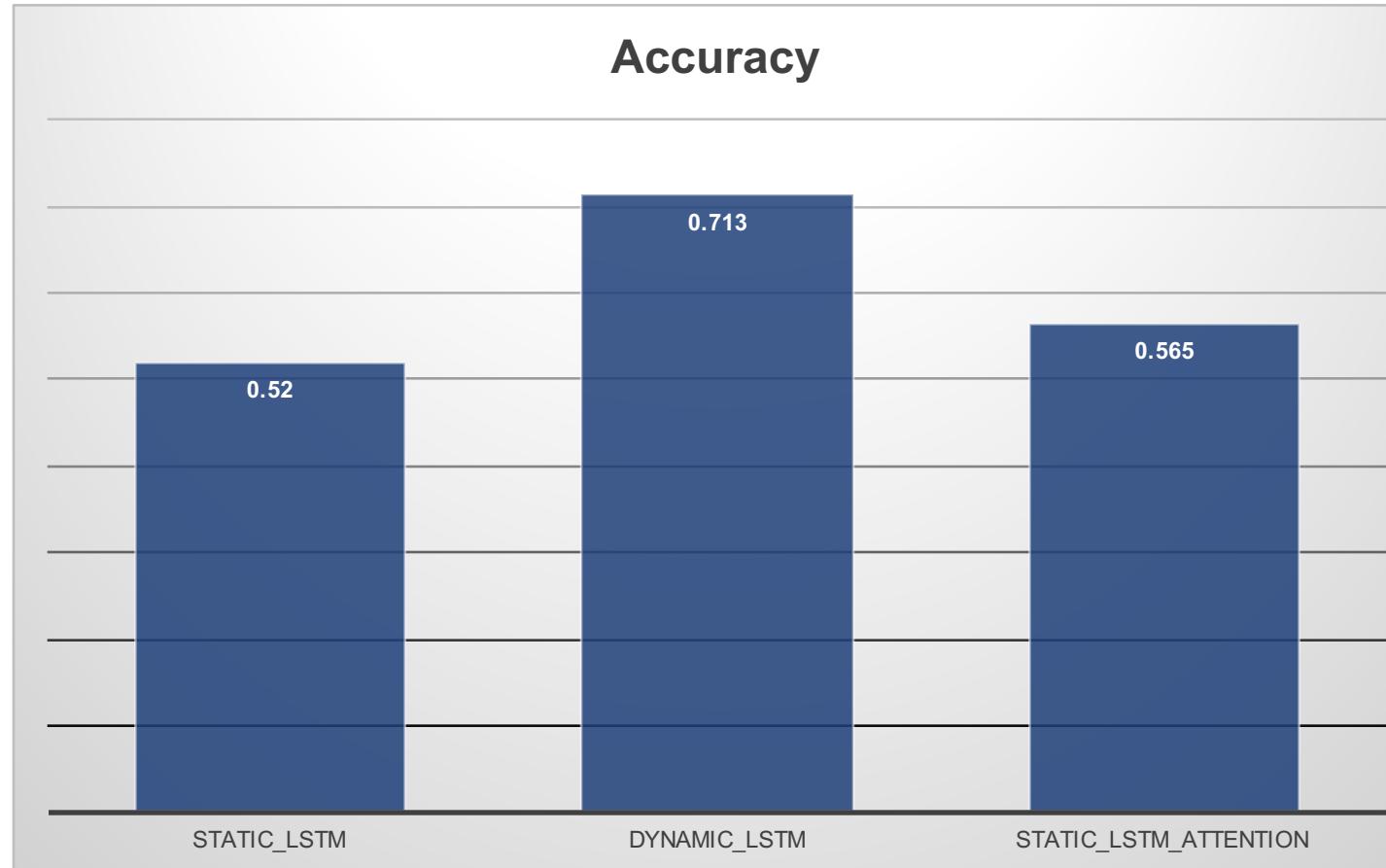
- Different Num filter: 64/128/256/512/1024



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.

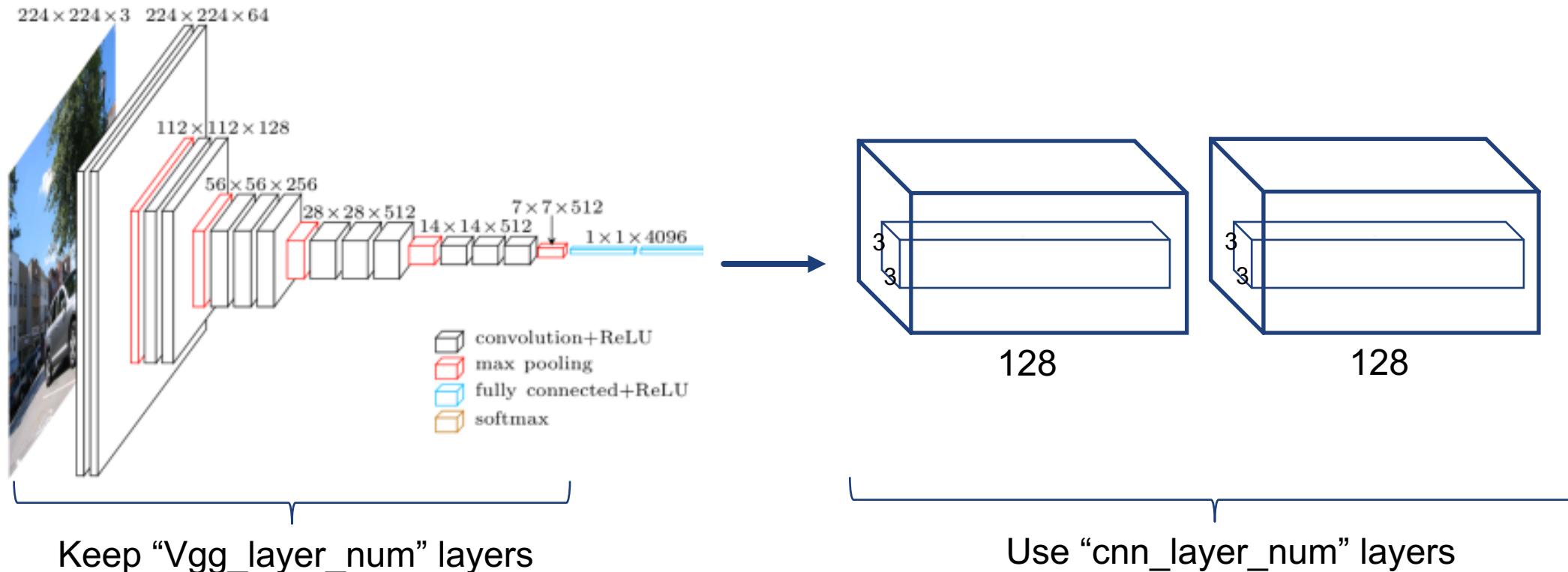
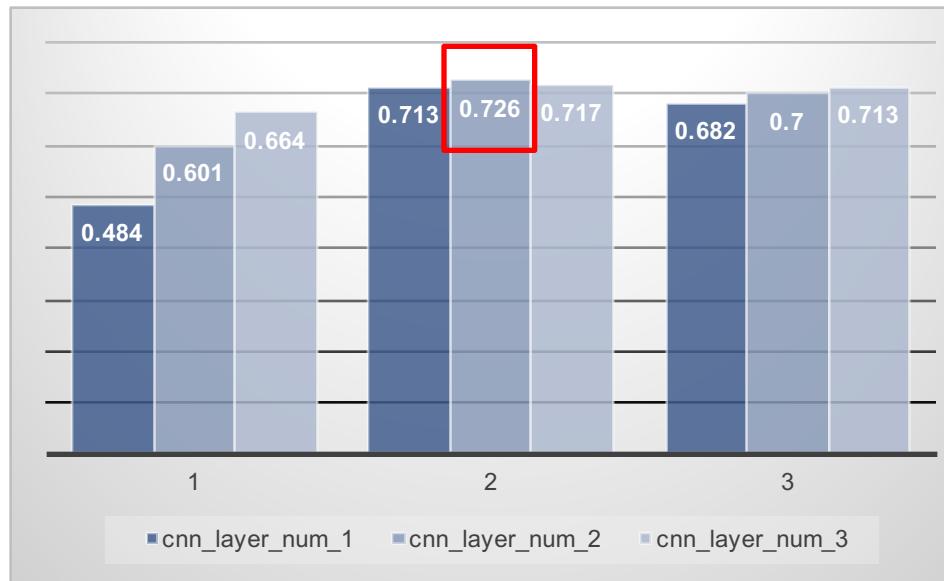
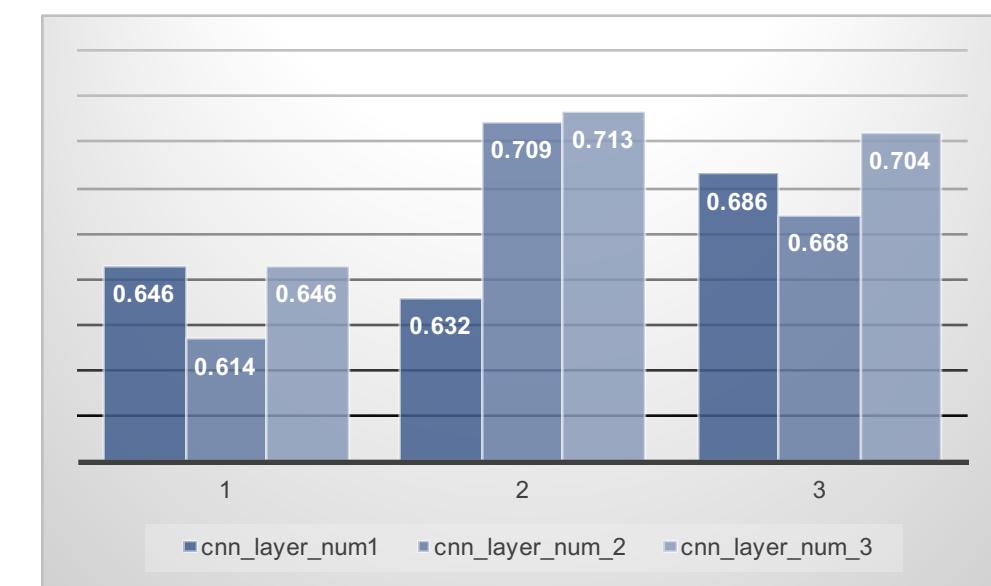


Image-CNN-Screenshot

- Training Time: 800 minutes



Trainable: False



Trainable: True

Compare screenshot and Layout

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

	True	4	16	500	224	3, 128, 3, 128, 3, 128	1, 1, 1	128
	True	4	16	500	224	3, 128, 3, 128, 3, 128	1, 1, 1	128
	True	4	16	500	224	3, 128, 3, 128, 3, 128	1, 1, 1	128
average accuracy: 0.726								
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

	True	4	16	500	224	224	224	3, 128, 3, 128, 3, 128	1, 1, 1	128
	True	4	16	500	224	224	224	3, 128, 3, 128, 3, 128	1, 1, 1	128
	True	4	16	500	224	224	224	3, 128, 3, 128, 3, 128	1, 1, 1	128
average accuracy: 0.682										
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	1,65	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

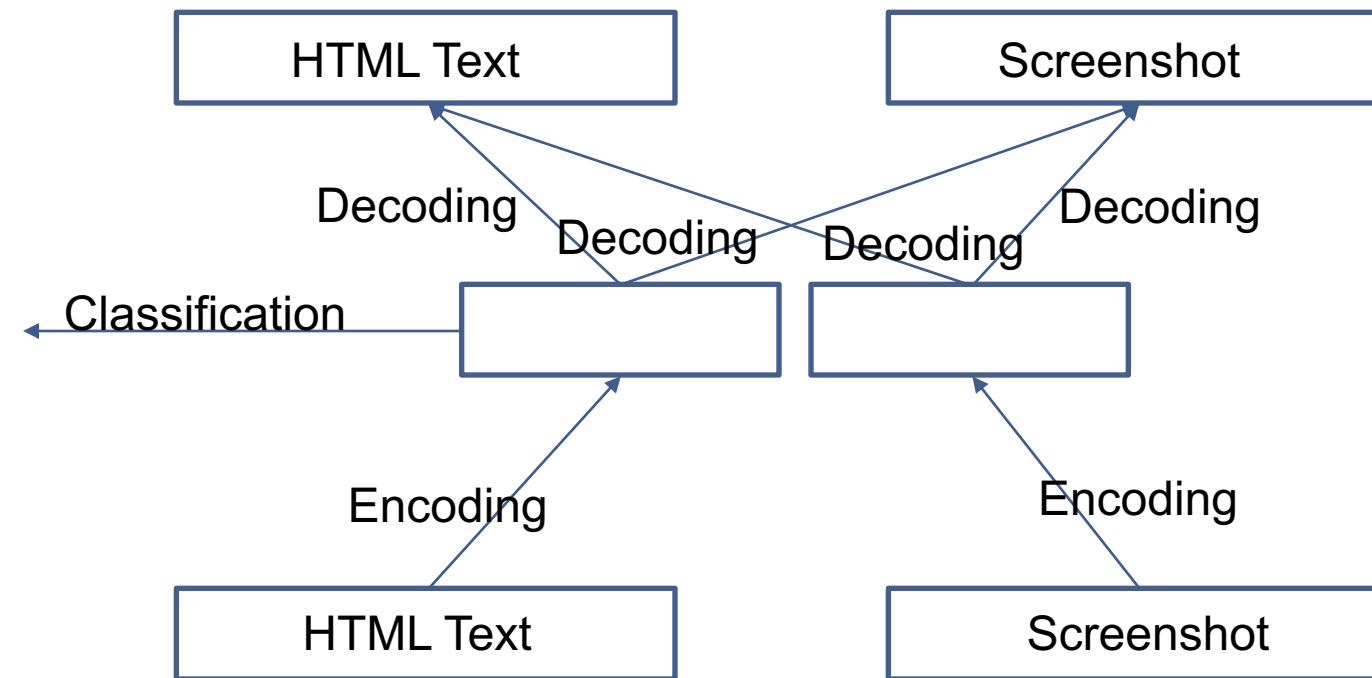
Problem and Future Work

- 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