Classification of Documents and Webpages

Supervised By: Dr. Fayyaz-ul-Amir Afsar Minhas
Presented By: Rao Muhammad Umer
Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan.

Outline

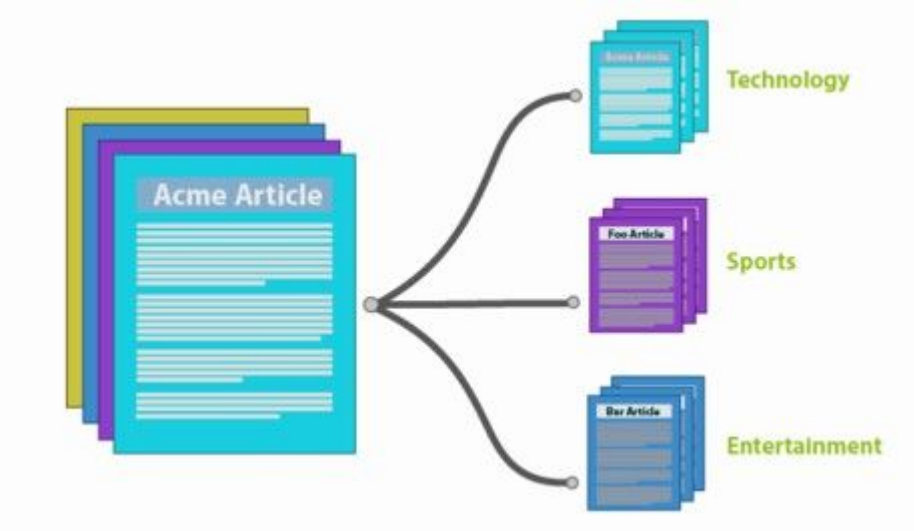
- Problem ?
- Is it a Machine Learning Problem ?
- Datasets ?
- Model Selection ?
- Performance Evaluation ?
- Conclusion ?
- Future Work ?

Outline

- **Problem ?**
- Is it a Machine Learning Problem ?
- Datasets ?
- Model Selection ?
- Performance Evaluation ?
- Conclusion ?
- Future Work ?

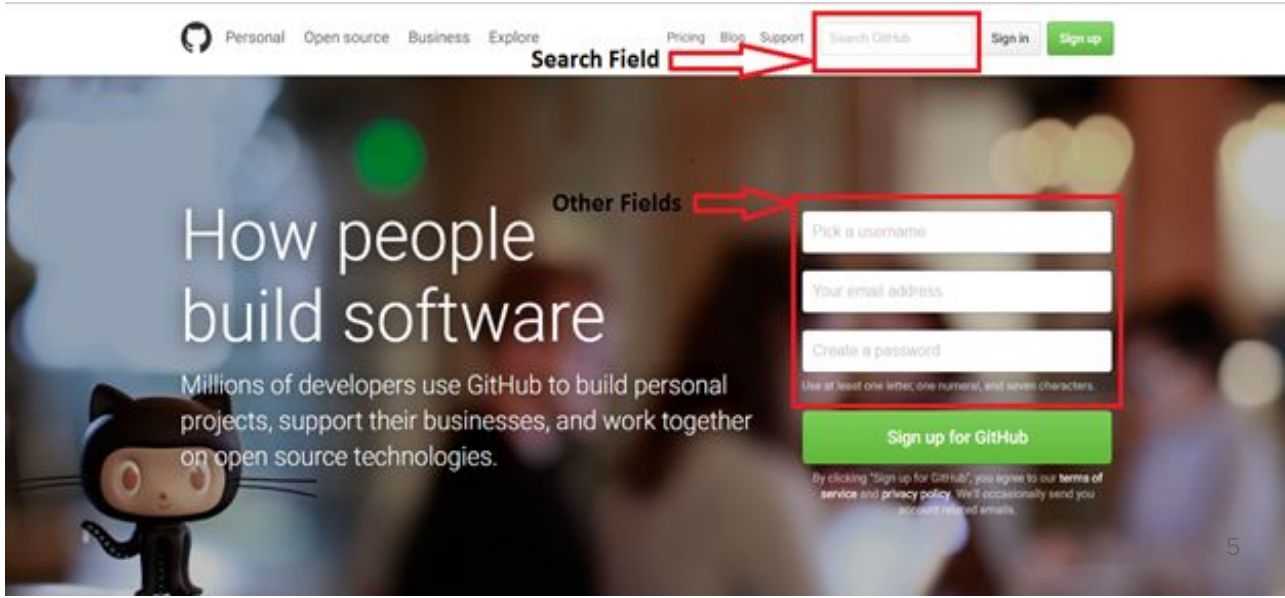
Documents Classification

- We want to predict the category/class label of a given test document (Text Classification).



Webpages Classification

- We want to classify the given web page as contain a searchable form or not.
- We also want to detect the fields of a given webpage.



https://github.com/

Problem Significance

- Text Categorization Techniques are used:
  - To classify news stories
  - To classify academic papers by technical domains and sub-domains
  - To classify patient reports in health-care organizations by using taxonomies of disease categories
  - To classify emails as spam or non-spam
  - etc.

Outline

- Problem ?
- **Is it a Machine Learning Problem ?**
- Datasets ?
- Model Selection ?
- Performance Evaluation ?
- Conclusion ?
- Future Work ?

Machine Learning Problem

- **Document Classification & Webpages Classification**
  - Are both ML problems?
  - Answer: **Yes**
- **How?**
  - Supervised Learning Problems
    - Classification
      - Predict the category of given document
      - Classify the searchable form page or not
    - Labeled Training Data
    - Output is also a label

Outline

- Problem ?
- Is it a Machine Learning Problem ?
- **Datasets ?**
- Model Selection ?
- Performance Evaluation ?
- Conclusion ?
- Future Work ?

Datasets

- **In Documents Classification task,**
  - **20-newsgroups dataset**

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

http://qwone.com/~jason/20Newsgroups/

- **In Webpages Classification task,**
  - **1000+ annotated web forms dataset**
  - https://github.com/RaoUmer/Formasaurus/tree/master/formasaurus/data

Outline

- Problem ?
- Is it a Machine Learning Problem ?
- Datasets ?
- **Model Selection ?**
- Performance Evaluation ?
- Conclusion ?
- Future Work ?

Model Selection

- **In Document Classification Task,**
  - Apply different classification models
    - **Multinomial Naive Bayes (MNB)**
    - **K-Nearest Neighbors (KNN)**
    - **Support Vector Machine (SVM)**
- **In Webpages Classification Task,**
  - Form type detection
    - **Support Vector Machine (Linear SVM)**
  - Field type detection
    - **CRFs (Conditional Random Fields)**

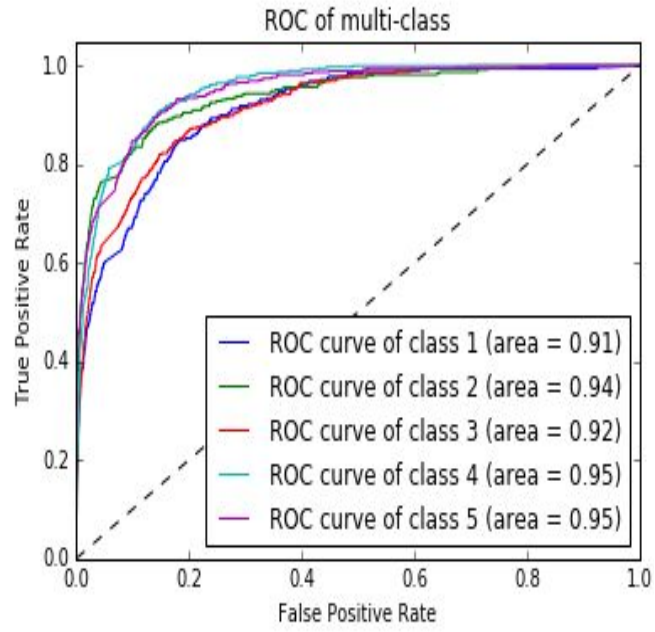
Outline

- Problem ?
- Is it a Machine Learning Problem ?
- Datasets ?
- Model Selection ?
- **Performance Evaluation ?**
- Conclusion ?
- Future Work ?

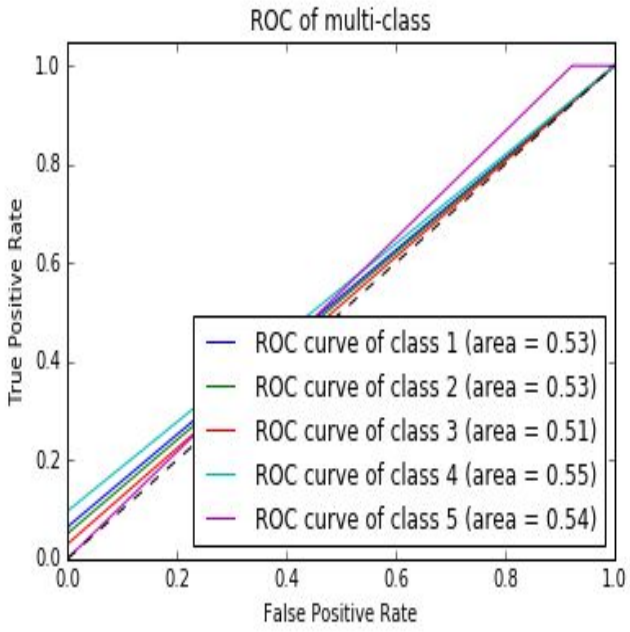
Documents Classification

- ROC and AUC

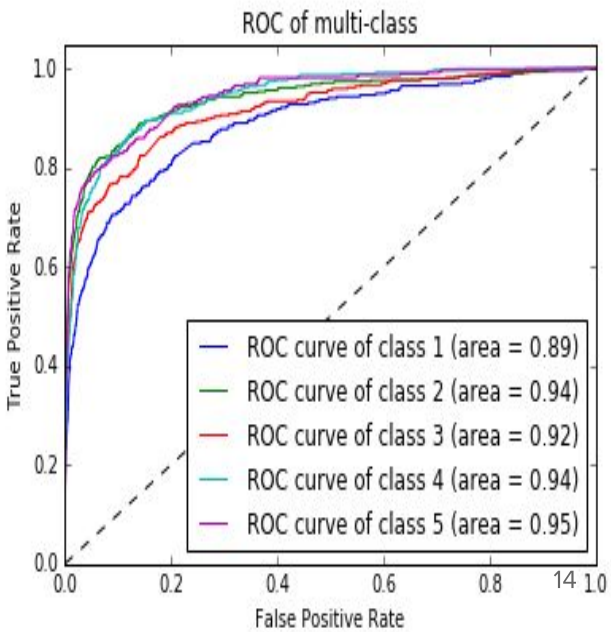
ROC and AUC of MultinomialNB



ROC and AUC of KNN



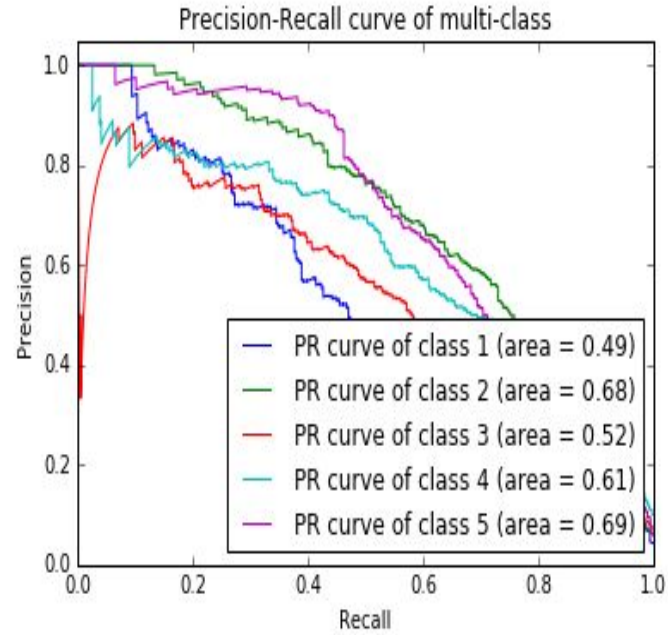
ROC and AUC of LinearSVC



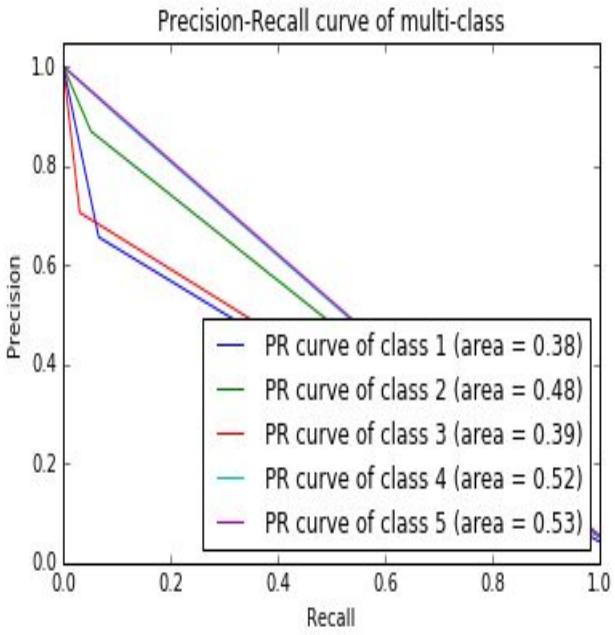
Documents Classification

- PR and AUC

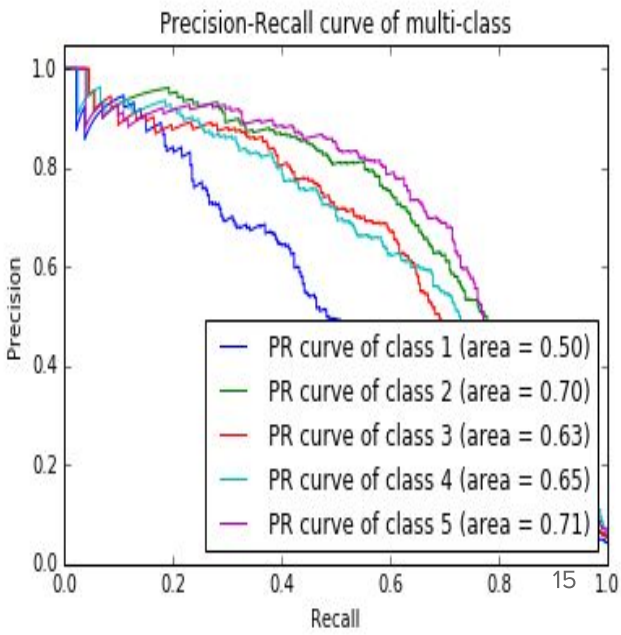
PR and AUC of MultinomialNB



PR and AUC of KNN



PR and AUC of LinearSVC



Documents Classification

- ROC & PR AUC Analysis

ROC-AUC Score						PR-ACU Score					
Multinomial Naïve Bayes		KNN		Linear SVM		Multinomial Naïve Bayes		KNN		Linear SVM	
Class No.	ROC-AUC	Class No.	ROC-AUC	Class No.	ROC-AUC	Class No.	PR-AUC	Class No.	PR-AUC	Class No.	PR-AUC
1	0.91	1	0.53	1	0.89	1	0.49	1	0.38	1	0.50
2	0.94	2	0.53	2	0.94	2	0.68	2	0.48	2	0.70
3	0.92	3	0.51	3	0.92	3	0.52	3	0.39	3	0.63
4	0.95	4	0.55	4	0.94	4	0.61	4	0.52	4	0.65
5	0.95	5	0.54	5	0.95	5	0.69	5	0.53	5	0.71
6	0.97	6	0.53	6	0.95	6	0.81	6	0.52	6	0.79
7	0.97	7	0.56	7	0.97	7	0.80	7	0.57	7	0.83
8	0.97	8	0.51	8	0.96	8	0.76	8	0.36	8	0.76
9	0.97	9	0.53	9	0.96	9	0.78	9	0.49	9	0.81
10	0.98	10	0.52	10	0.97	10	0.87	10	0.42	10	0.86
11	0.99	11	0.55	11	0.99	11	0.94	11	0.57	11	0.91
12	0.97	12	0.53	12	0.95	12	0.81	12	0.41	12	0.79
13	0.91	13	0.52	13	0.90	13	0.54	13	0.42	13	0.58
14	0.97	14	0.52	14	0.95	14	0.81	14	0.51	14	0.81
15	0.97	15	0.53	15	0.96	15	0.79	15	0.49	15	0.79
16	0.97	16	0.54	16	0.96	16	0.76	16	0.49	16	0.73
17	0.95	17	0.51	17	0.93	17	0.55	17	0.32	17	0.57
18	0.98	18	0.58	18	0.97	18	0.87	18	0.56	18	0.85
19	0.88	19	0.53	19	0.83	19	0.49	19	0.55	19	0.46
20	0.89	20	0.52	20	0.87	20	0.27	20	0.41	20	0.31

Webpages Classification

- **Useful features for Searchable Form page**
  - a single query field
  - a field named **"q"** or **"s"**
  - **"search"** in URL
  - **"search"** in submit button text (submit value)
  - **"search"** in form css class or id
  - no password field
  - method == **GET/POST**
  - etc.

Webpages Classification

- **Useful features for Field type detection of webpage**
  - form type predicted by a form type detector
  - field tag name
  - field value
  - text before and after field
  - field CSS class and ID
  - text of field label element
  - field title and placeholder attributes
  - etc.

Webpages Classification

- Form Type Detection

Annotated HTML forms (simplified classes)

415	search	(s)
246	login	(l)
165	registration	(r)
143	other	(o)
138	contact/comment	(c)
132	join mailing list	(m)
105	password/login recovery	(p)
74	order/add to cart	(b)

Total form count: 1418

	precision	recall	f1-score	support
search	0.92	0.96	0.94	415
login	0.96	0.96	0.96	246
registration	0.95	0.87	0.91	165
password/login recovery	0.86	0.84	0.85	105
contact/comment	0.85	0.94	0.89	138
join mailing list	0.88	0.88	0.88	132
order/add to cart	0.96	0.62	0.75	74
other	0.66	0.71	0.68	143
avg / total	0.89	0.89	0.89	1418

88.7% forms are classified correctly.

Webpages Classification

- Field Type Detection

	precision	recall	f1-score	support
search query	0.843	0.980	0.907	99
email	0.945	0.987	0.966	156
password	1.000	0.966	0.983	88
product quantity	1.000	0.875	0.933	8
submit button	0.895	1.000	0.944	68
username	0.767	0.767	0.767	43
password confirmation	1.000	1.000	1.000	24
receive emails confirmation	0.909	0.370	0.526	27
first name	0.913	0.840	0.875	25
last name	0.870	0.800	0.833	25
organization name	1.000	0.417	0.588	12
address	0.706	0.667	0.686	18
city	0.909	0.714	0.800	14
state	1.000	0.750	0.857	4
postal code	1.000	0.929	0.963	14
country	0.875	0.636	0.737	11
phone	1.000	0.944	0.971	18
fax	1.000	1.000	1.000	1
TOS confirmation	1.000	0.692	0.818	13
comment text	0.786	0.971	0.868	34
captcha	0.962	0.735	0.833	34
remember me checkbox	1.000	1.000	1.000	29
username or email	0.667	0.222	0.333	9
other	0.730	0.854	0.787	171
full name	0.595	0.926	0.725	27
search category / refinement	0.842	0.985	0.908	65

Webpages Classification

- Testing on web pages that didn’t give in training phase for generalization

Webpages Classification

https://github.com/

Webpages Classification

- **Results (TP)**

[(<Element form at 0x89bbd18>, {**'fields':** {'q': 'search query'}, **'form':** u'search'}), (<Element form at 0x89bbd68>, {**'fields':** {'user[password]': 'password', 'user[login]': 'username', 'user[email]': 'email'}, **'form':** u'registration'})]

Webpages Classification

https://sites.google.com/site/drmabidm/

Webpages Classification

- **Results (TP)**

[(<Element form at 0xb0fa318>, {**'fields':** {'q': 'search query'}, **'form':** u'search'})]

Webpages Classification

http://www.pieas.edu.pk/

Webpages Classification

- **Results (TN)**

[NULL]

Webpages Classification

Fayyaz-ul-Amir Afsar Minhas ¶

Hello and Welcome to my research and teaching profile! I am a senior scientist at the Department of Computer and Information Sciences (DCIS) at Pakistan Institute of Engineering and Applied Sciences (PIEAS), Pakistan. I am involved in active research and teaching in computer science. Here, you can access information about my research projects and academic courses.

I graduated with my Ph.D. in Computer Science from Colorado State University (Go Rams!), Fort Collins, Colorado, USA under a Fulbright scholarship with Dr. Asa Ben-Hur. My primary area of research is machine learning in Bioinformatics. You can view details of my research lab here.

I am currently looking for graduate (MS/MPhil and Ph.D.) students in computer science to work on problems in the lab. If you are interested in research and development in these or related areas, please feel free to contact me.

You can also do a custom search of this website using the search box below.

Google™ Custom Search

Search field

Contact

Dr. Fayyaz-ul-Amir Afsar Minhas
Senior Scientist
Department of Computer & Information Sciences (DCIS)
Pakistan Institute of Engineering & Applied Sciences (PIEAS)
PO Nilore, Islamabad, Pakistan 45650

Office: B-Block, Room 216
Phone: +92-51-2207381 to 85, Ext. 3164
Email: fayyazafsar <at> gmail [.] com, afsar <at> pieas [.] edu <dot> pk
Web: http://faculty.pieas.edu.pk/fayyaz

http://faculty.pieas.edu.pk/fayyaz/

Webpages Classification

- **Results (FN)**

[NULL]

Outline

- Problem ?
- Is it a Machine Learning Problem ?
- Datasets ?
- Model Selection ?
- Performance Evaluation ?
- **Conclusion ?**
- Future Work ?

Conclusion

- **In Documents classification task,**
  - Multinomial NB and Linear SVM performs quite well, while KNN fails on high dimensional feature space.
  - Multinomial NB and Linear SVM both have quite same ROC curve, but Linear SVM has good PR curve than that of Multinomial NB, While KNN has both ROC and PR curve as a random classifier.
- **In Webpages classification task,**
  - Linear SVM classifier has good PR and f1 measure score than that of other classifiers.
  - CRFs are good choice for sequence order of field type detection.

Outline

- Problem ?
- Is it a Machine Learning Problem ?
- Datasets ?
- Model Selection ?
- Performance Evaluation ?
- Conclusion ?
- **Future Work ?**

Future Work

- **In documents classification,**
  - Since in above mentioned all classification models have assumption that **features are independent** of each other (use Bag of words model), but in real scenario, the situation is entirely different. Features are dependent on each other and context of sentence.
  - So, documents classification task would be better accurate results if it would be consider as a **NLP problem** (dependencies among features).
  - In 20-newsgroups dataset, we only tackled the 20 classes of multi-class classification task by **OVR (One-Vs-Rest)** classifier that would also lower our accuracy of model.
  - In 20-newsgroups dataset, in each category, there are also sub-categories in each class. In our model, we only tackled only single as root category. It would be better to predict the whole hierarchy of these categories that will lead to this problem to **Structured Learning** problem.

Future Work

- **In Webpages classification,**
  - Same **feature independence** assumption lies as in documents classification
  - To tackle them as **NLP problem**
  - To tackle the **Java-scripted based web-forms**

References

- http://www.scholarpedia.org/article/Text_categorization
- http://scikit-learn.org/stable/
- http://scikitlearn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
- http://scikitlearn.org/stable/auto_examples/model_selection/plot_roc.html#example-model-selection-plotroc-py
- http://scikitlearn.org/stable/model_selection.html
- https://github.com/TeamHG-Memex/Formasaurus.git
- http://formasaurus.readthedocs.org/en/latest/