



# Internet Archive PDF

Huilin Chang

# Presentation Outline

- Problem Statement
- Data Overview
- Data Pipeline
- Data Feature Engineering
- Exploratory Data Analysis
- Models
- Conclusion

# Research paper example

*ECS Journal of Solid State Science and Technology*, **3** (10) M65-M70 (2014) M65

## Synthesis and Characterization of Silicon and Nitrogen Containing Carbon-Based Crystals and Their Nanostructured Materials

**Hui Lin Chang,[a,b,z] Chi Tso Chang,[a] and Cheng Tzu. Kuo[a]**

[a]*National Chiao Tung University, Department of Material Science and Engineering, HsinChu, Taiwan*
[b]*Process Development Team (2PJT), Semiconductor R & D Center, Samsung Electronics, Korea*

The synthesis of carbon-based materials, such as man-made diamonds, superhard $C_3N_4$ materials, SiCN crystals, and other carbon-based nanostructured materials, has attracted considerable attention for many decades in academic and industrial communities. However, so far, researchers have not successfully linked the growth mechanisms of carbon-based materials deposited under different synthetic conditions and methods. In fact, a single machine may produce many of these materials. This paper is aimed to study the linkages among various carbon-based materials synthesized on Si wafers using the same microwave plasma chemical vapor deposition system, including SiCN crystalline films, SiCN nanotubes, carbon nanotubes (CNTs), conical carbon nanorods, and other nanostructured materials.



Manuscript submitted July 3, 2014; revised manuscript received July 28, 2014. Published August 21, 2014.

Sustaining Moore's law requires constant transistor scaling, boosting the creation of new materials for future nanoelectronics applications. Several emerging materials, such as Si nanowires, carbon nanotubes (CNTs), and III–V semiconductor field effect transistors (FETs), are potential components in this continuous shrinking process. In particular, CNTs are expected to overcome the physical limitation of current Si transistors and Cu interconnections in molecular electronics.[1–5] However, their integration into Si-based metal-oxide-semiconductor field effect transistors (MOSFETs) or new nanoelectronics remains challenging when developing these transistors and interconnections. They are naturally deposited as bundles in a vertical direction because they tend to adhere to each other vertically. Vertically aligned nanotube field-effect transistors (VCNTFETs) have been proposed to yield the Si device characteristics required for 2016, as set by the International Technology Roadmap for Semiconductors.[6–12] The feasibility of this vision depends on direct approaches to achieve selective depositions in the trenches or holes of Si wafers. The deposition of CNTs bundles in the trenches and holes as channels and conductors, respectively, can provide sufficient current density. The manipulation of CNTs orientation in either horizontal or vertical direction also plays a key role in manufacturing. This study systematically evaluates the synthesis of CNTs by microwave plasma chemical vapor deposition (MPCVD) using an Fe catalyst, a $CoSi_2$ film, and Ni islands, which frequently serve as gate electrodes and contact materials in Si microelectronics. The selective growth of CNTs in trench/hole/planar forms is also examined in conjunction with their morphology and nanostructures. The field emission characteristics of CNTs deposited in trenches and holes are examined to determine electronic performance. Moreover, the electronics properties of nanocrystals and tubular structures are compared. The growth mechanism and electronic properties of nanostructured materials are addressed.

### Experimental

Figure 1 schematically shows relative positions between plasma and sample in a MPCVD system. Figure 2 compares various carbon-based materials synthesized on Si wafers using the same MPCVD system. Process parameters are divided into three groups, i.e., nanotubes, nanowires, and nanocrystals, according to the synthesized material structures. Main parameters include temperature, reactive gas type ($CH_4/H_2$; $CH_4/N_2$; $CH_4/H_2/N_2$), catalyst (Fe, Co Ni), additional Si source, and patterning design for selective CNTs growth. Deposition temperatures are estimated by placing a thermocouple under a substrate holder. Nanotube morphologies and microstructures were identified by scanning (SEM) and transmission (TEM) electron microscopes. Field emission properties were evaluated by I-V measurements at $10^{-6}$ Torr for electrode separations of 50 and 100 μm. Table I lists the detail parameters of each sample and its corresponding morphology.

### Results and Discussion

*Nanostructured material synthesis by MPCVD.*— Figure 2 shows the linkages among various carbon-based materials synthesized on Si wafers using the same MPCVD system at different parameters, such as temperature, gas types ($CH_4/H_2$, $N_2$), deposited buffer layer (Co and Ti),[13] additional Si source, and pattern design for selective CNTs growth. Three routes (①, ② and ③ in Fig. 2) were compared for the catalyst-assisted synthesis of carbon nanowires and nanorods as well as selective CNTs growth. The results reveal that formation and properties of CNTs can be manipulated by applying catalysts with $H_2$ reduction gas ($CH_4/H_2$ ratio = 10 sccm /100 sccm, temperature at 500°C) leading to CNTs formation (Fig. 3). In contrast, Condition 2 (route ② in Fig. 2) is under lower $CH_4/H_2$ ratio ($CH_4/H_2$ ratio = 1 sccm /100 sccm, temperature at 450°C) leading to nanowire formation. (Fig. 4). The $CH_4/H_2$ ratio influences the formation of tubular and crystalline structures. A high $CH_4/H_2$ ratio favors the formation of C-$sp^2$ bonding (graphite structure), whereas, a low $CH_4/H_2$ ratio favors the formation of C-$sp^3$ bonding (diamond structure). Therefore, carbon atoms surround the catalysts and later precipitate from them with different $CH_4/H_2$ ratios to form hollow tubes or solid nano-wires. Under Condition 3 (route ③ in Fig. 2), CNTs were selectively deposited on patterned wafers, such as (a) parallel Fe-coated line arrays and (b) $CoSi_2$-coated hole arrays. This novel method is compatible with Si microelectronic device manufacturing, as shown in Figs. 5 and 6. In addition, Fig. 5b shows 18 μm-long CNTs selectively deposited on Fe-coated line arrays at 3 μm/min. These CNTs are essentially well aligned, uniform in size, and perpendicular to the substrate. Fig. 6 shows that CNTs are also selectively deposited in the holes of Si wafers patterned with hole arrays (aspect ratio 6). SEM micrographs reveal the 5 nm-diameter CNTs are wrapped inside the holes rather than forming well-aligned CNTs (Fig. 5) under similar conditions, suggesting the high selectivity of this process. The wrapping of the CNTs in the holes may result from the local circular flow of gases in each hole.

To summarize, nanostructured carbon nanowires and nanorods were successfully synthesized on patterned and unpatterned Si wafers in the presence of a catalyst by varying the process parameters, such as catalyst materials, source gases, gas ratios, and deposition temperatures. This result also offers a different perspective on the mechanism of the catalyst-assisted MWCNTs growth. The $CH_4/H_2$ ratio influences the formation of tubular and crystalline structures. Specifically,

[z]E-mail: huilinchang@fas.harvard.edu



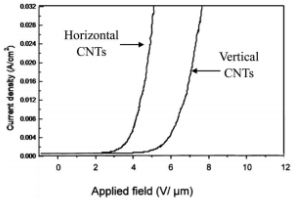



**Figure 15.** J-E curves of CNTs with preference orientation horizontal and vertical to the substrate. The corresponding current density of 1mA/cm$^2$ is obtained at 1.57 and 4.32 V/μm, respectively.

### Conclusions

On account of Moore's law which relies on constant transistor scaling, emerging materials such as Si nanowires, carbon nanotubes, and III–V semiconductor FETs, are expected to transform future nanoelectronics applications validating this theory. This study demonstrates selective CNTs deposition methods that lead to vertically and horizontally oriented growth. In addition, several SiCN and carbon-based nanostructures are successfully synthesized using the same MPCVD system. The development of nanostructured materials with unique electrical properties may expand nanoelectronic device applications.

### Acknowledgments

The authors thank the National Science Council of the Republic of China, Taiwan for financially supporting this research. The authors also thank Dr. M. S. Liang, Dr. E. S. Jung and Dr. J. H. Ku, Dr. M. C. Kim and Dr. B. Y. Ku for their valuable discussion.

### References

1. V. V. Mitin, V. A. Kochelap, and M. A. Stroscio, *Introduction to nanoelectronics* (Cambridge, 2008), p10.
2. S. Iijima, *Nature (Lond.)* **354**, 56 (1991).
3. T. W. Ebbesen, H. J. Lezec, H. Hiura, J. W. Bennett, H. F. Ghaemi, and T. Thio, *Nature* **382**, 54 (1996).
4. W. A. de Heer, W. A. Chatelain, and D. A. Ugarte, *Science* **270**, 1179 (1995).
5. G. S. Duesberg, A. P. Graham, F. Kreupl, M. Liebau, R. Seidel, E. Unger, and W. Hoenlein, *Dia. Relat. Mat.* **13**, 354 (2004).
6. B. Q. Wei, R. Vajtai, and P. M. Ajayan, *Appl. Phys. Lett.* **79**, 1172 (2001).
7. P. Avouris, *Acc. Chem. Res.* **35**, 1026 (2002).
8. V. Derycke, R. Martel, J. Appenzeller, and P. Avouris, *Nano Lett.* **1**, 453 (2001).
9. J. Chaste, L. Lechner, P. Morfin, G. Fve, T. Kontos, J. M. Berroir, D. C. Glattli, H. Happy, P. Hakonen, and B. Plaais, *Nano Lett.* **8**, 525 (2009).
10. R. V. Seidel, A. P. Graham, J. Kretz, B. Rajasekharan, G. S. Duesberg, M. Liebau, E. Unger, F. Kreupl, and W. Hoenlein, *Nano Lett.* **5**, 147 (2005).
11. Sander J. Tans, Alwin R. M. Verchueren, and Cees Dekker, *Nature* **393**, 49 (1998).
12. W. Hoenlein, F. Kreupl, G. S. Duesberg, A. P. Graham, M. Liebau, R. Seidel, and E. Unger, *Mater. Sci. & Eng. C* **23**, 663 (2003).
13. H. L. Chang and C. T. Kuo, *Diamond Relat. Mater.* **10**, 1910 (2001).
14. Y. Saito and T. Yoshikawa, *J. Cryst. Growth* **134**, 154 (1993).
15. H. Murakami, M. Hirakawa, C. Tanaka, and H. Yamakawa, *Appl. Phys. Lett.* **76**, 1776 (2000).
16. D. C. Li, L. Dai, S. Huang, A. W. H. Mau, and Z. L. Wang, *Chem. Phys. Lett.* **316**, 349 (2000).
17. D. Zhou and S. Seraphin, *Chem. Phys. Lett.* **238**, 286 (1995).
18. M. Endo, S. Iijima, and M. S. Dresselhays(Eds.) *Carbon Nanotubes*, BPC Press, UK, pp. 37, 159 (1999).
19. C. H. Lin, H. L. Chang, C. M. Jsu, A. Y. Lo, and C. T. Kuo, *Diamond Relat. Mater.* **12**, 1851 (2003).
20. H. L. Chang, J. S. Fung, and C. T. Kuo, *Rev. Adv. Mater. Sci.* **5**, 432 (2003).
21. M. K. Sunkara, S. Shuma, and R. Miranda, *Appl. Phys. Lett.* **79**, 1546 (2001).
22. C. A. Spindt, I. Brodie, I. Humphery, and E. R. Westerberg, *J. Appl. Phys.* **47**, 5248 (1976).

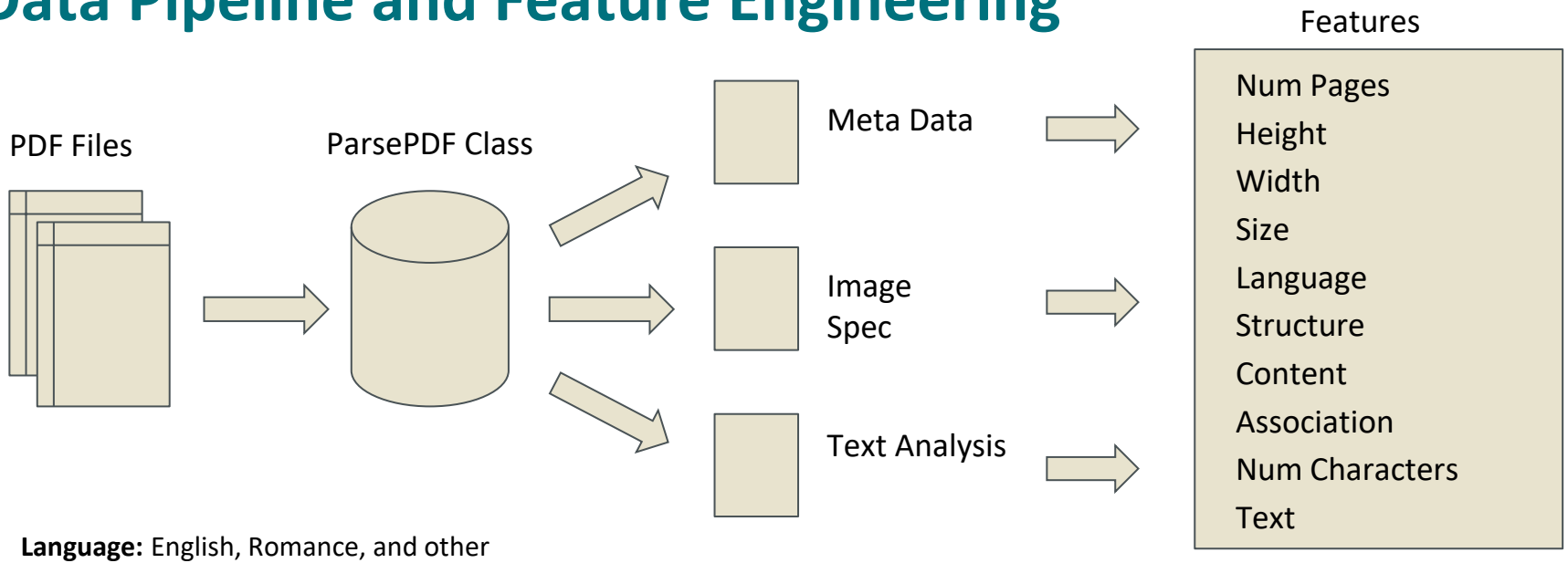
# Problem Statement

- One of the Internet Archive's mission areas is "Universal Access to All Knowledge", which includes collecting and providing access to the "scholarly web" -- research publications and datasets
- Curation to accurately identify legitimate research publications is needed to help users find scholarly content
- An inclusive approach that accounts for diverse content, particularly from underrepresented geographic areas, groups, and content domains, is important to avoid excluding relevant content due to implicit bias and narrow criteria
- Our project aims to help this mission by implementing a fast PDF identification tool, which will score files on their likelihood of being a research publication

# Data Overview

- 4 IA Training Datasets
  - Global Wayback Random - Random sampled PDFs from the Wayback Machine
  - Fatcat - A set of PDFs from the existing ‘Fatcat’ catalog of research papers
  - Fatcat Longtail Language - Papers from less-represented languages
  - Longtail - A set of PDFs created using heuristics (GROBID)
- Minor issues with data
  - Encrypted PDFs
  - Corrupted/Unparsable
- Plan to branch out further into IA content archives as well as other known sources of PDF scholarly documents

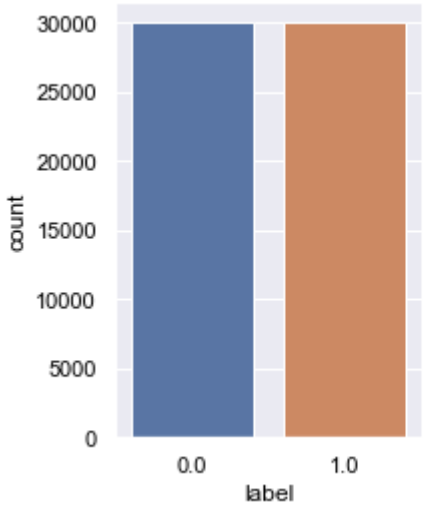
# Data Pipeline and Feature Engineering

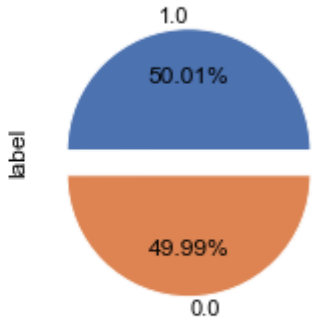


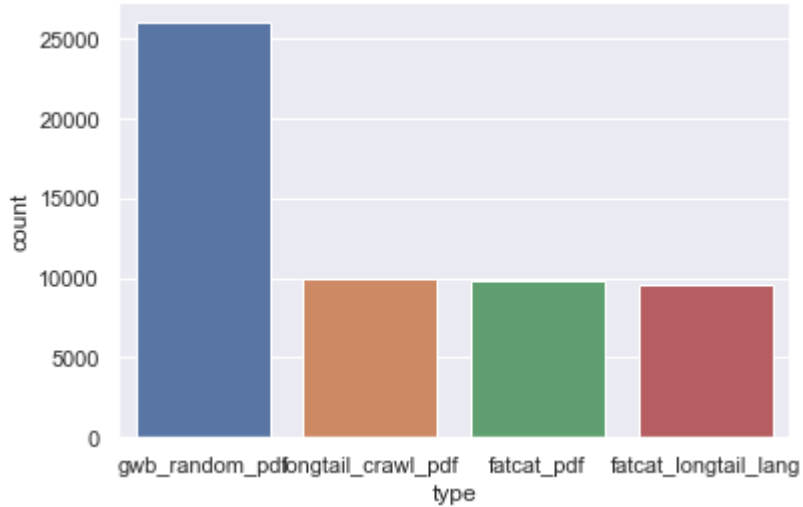

**Language:** English, Romance, and other
**Structure**: Words that represent the structure of a paper
{abstract, introduction, conclusion, reference, table of content}
**Content**: Words that represent the content of a paper
{research, analyze, result, table, investigation, explain, theory, study, paper, data, perform}
**Association**: Words that represent association
{journal, association, organization, doi, university, school, board}

# Balanced Data

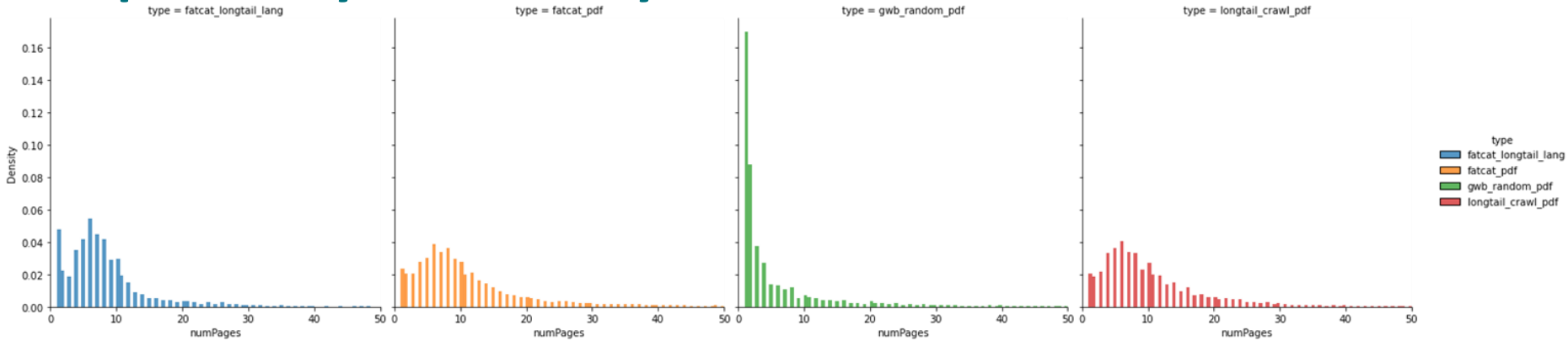
- Balanced dataset

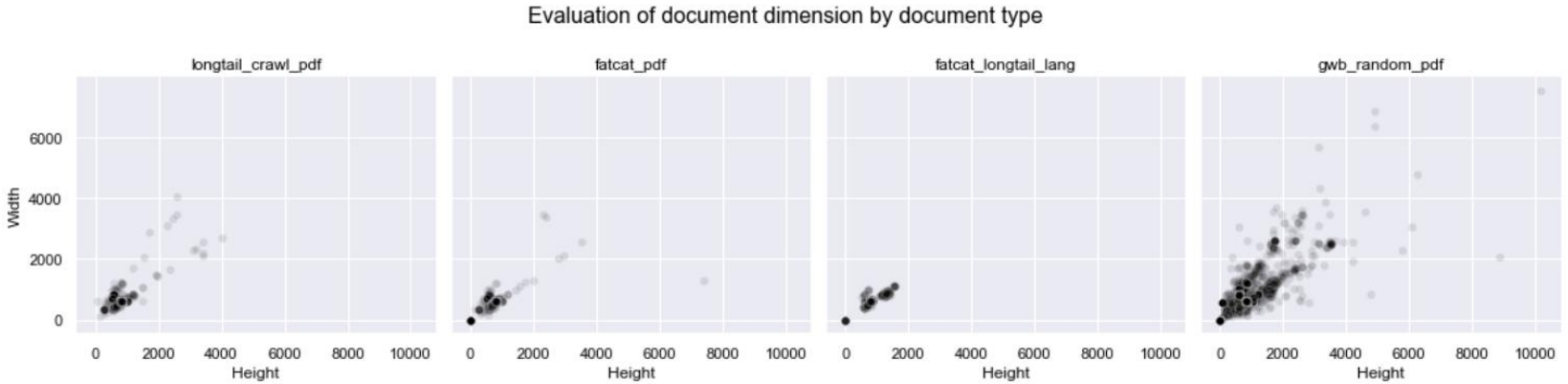


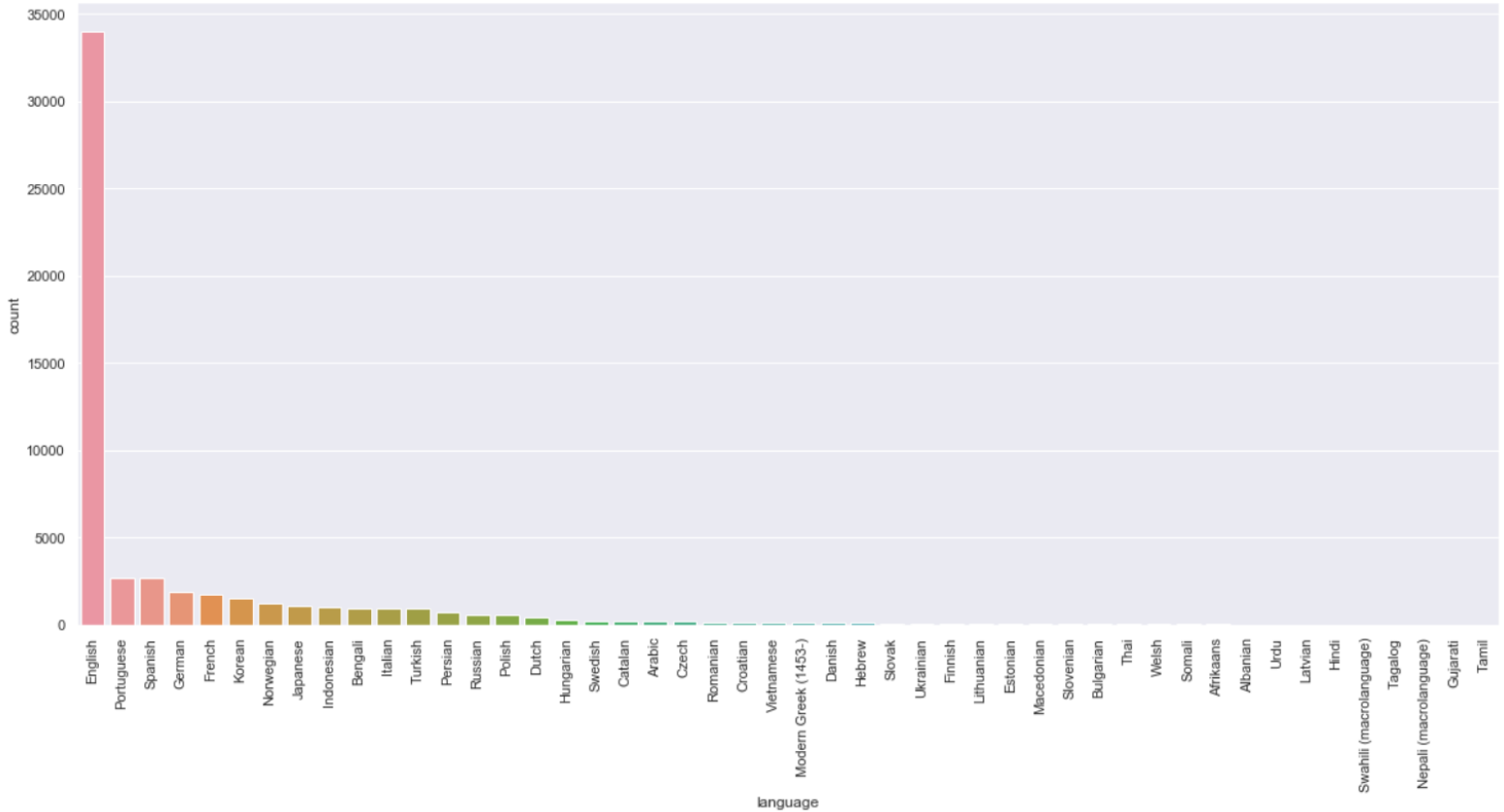



# Exploratory Data Analysis

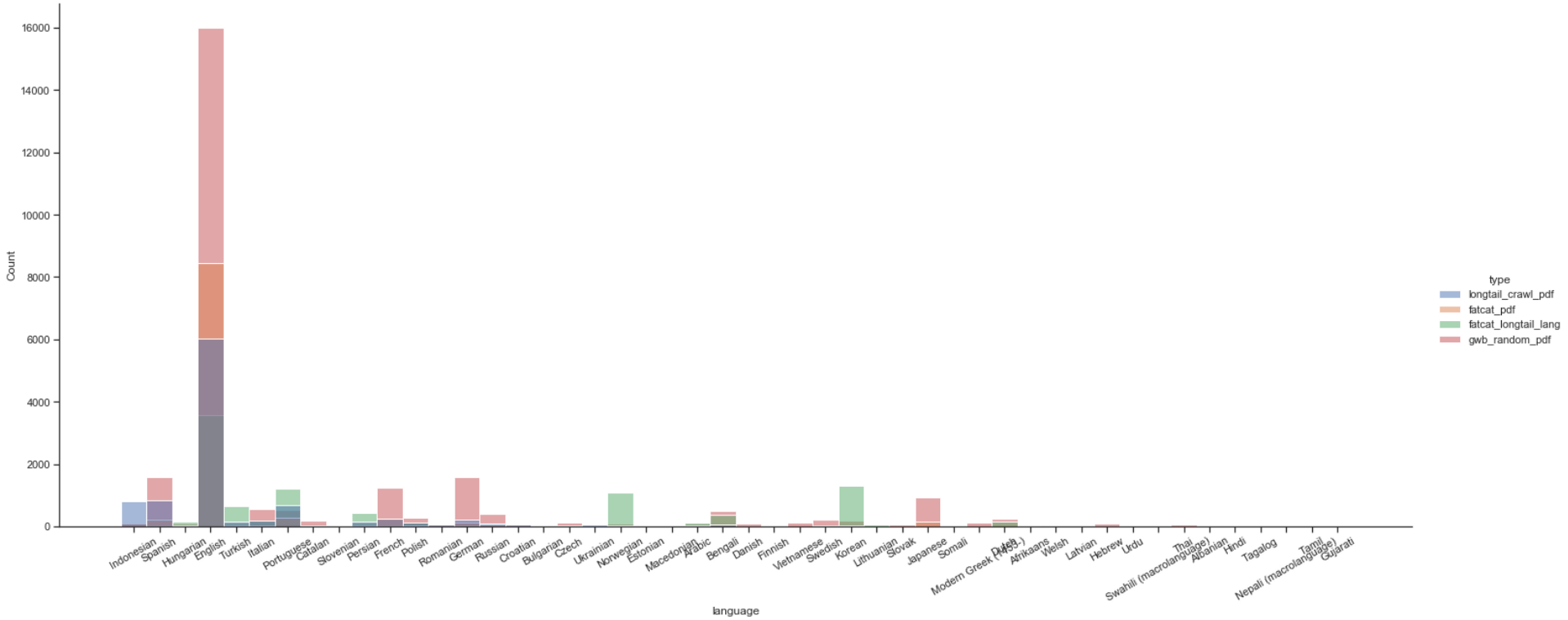


# Exploratory Data Analysis

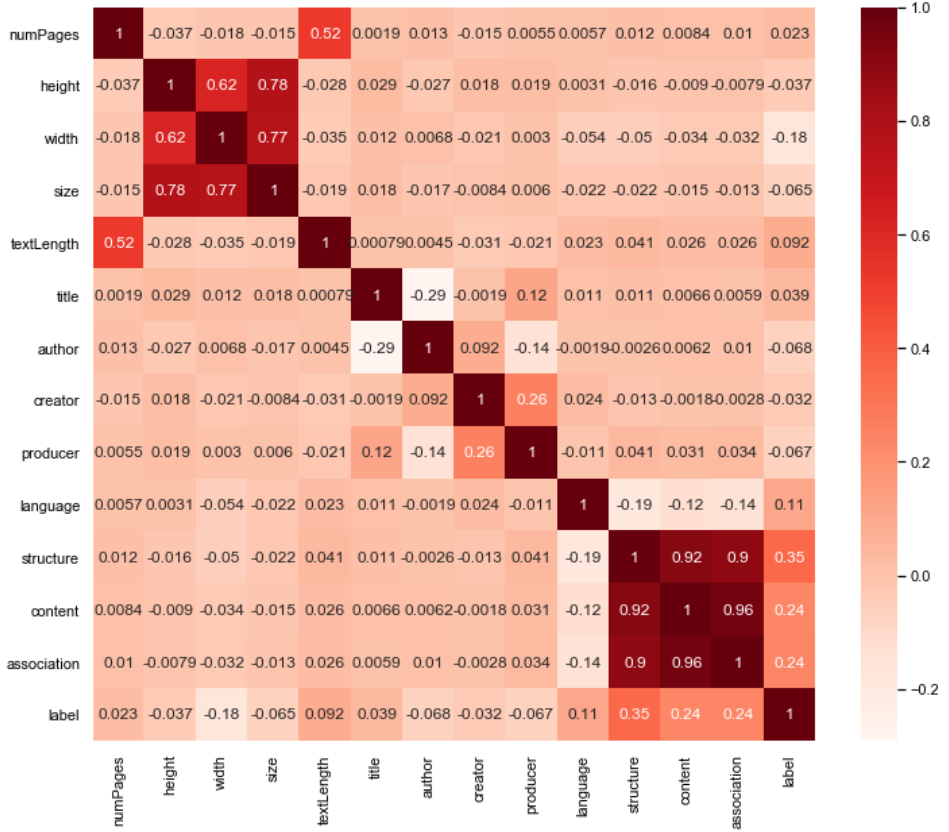


# Exploratory Data Analysis



# Data Feature Engineering

- The use of multiprocessing allows for further feature extraction
  - Ability to look for keywords in text
    - English
    - Non-english
      - Translate keywords to the language of the text
  - Process adds 14 minutes to the additionally extraction of meta and text data



# Models

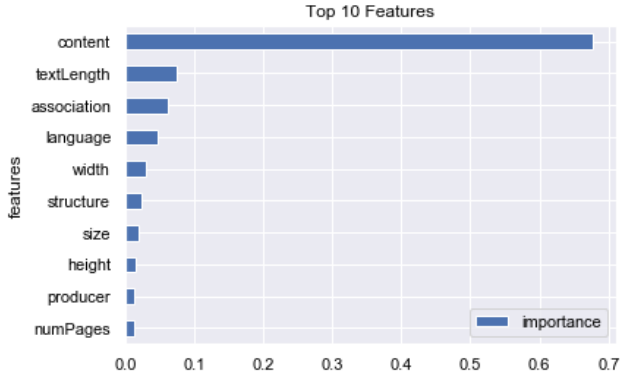
- Balanced Data
- Data Pipeline
- Data Feature Engineering
- Models
  - Text Based Models
    - XGBoost
    - Keras
    - SVM
  - Image Based Model
    - Keras (VGG16)
  - Bayesian statistic
    - Logistic models
  - Topic modeling (LDA)

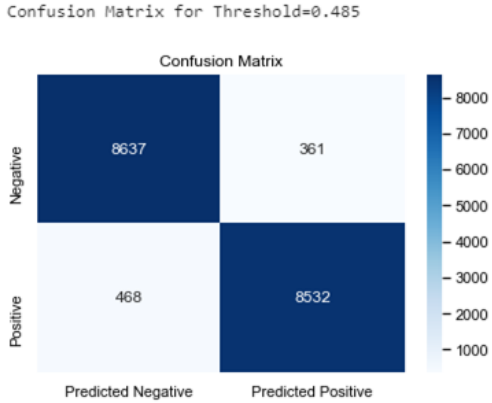
# Text Based Models

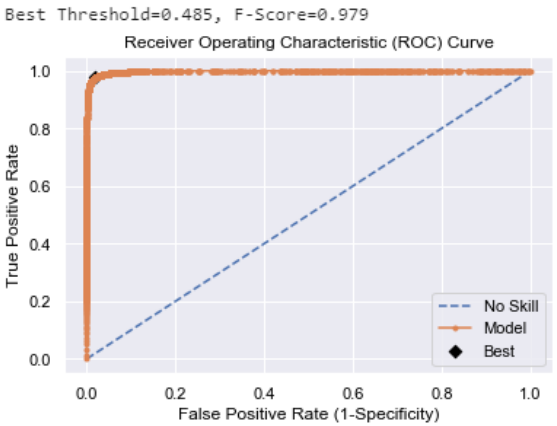
| Model Type | Accuracy |
| --- | --- |
| XGBoost | 95.39% |
| Keras | 93.89% |
| SVM | 90.40% |

# XGBoost

- Grid Search
  - N Estimators, Learning Rate, Depth
  - 3 k-fold
- F-score : 97.90%
- Accuracy:  95.39%



# Keras-Tensorflow

- Model Structure
  - Input dimension of 14
  - One hidden layer
  - Adam optimizer
  - Epochs 100
- Accuracy: 93.89%

```
from keras.models import Sequential
from keras.layers import Dense
import tensorflow as tf

model = Sequential()

model.add(Dense(2048, activation='relu', input_shape=(14,)))
model.add(Dense(1024, activation='relu', ))
opt = keras.optimizers.Adam(learning_rate = 0.001)

model.add(Dense(1, activation='sigmoid'))
```

```
model.compile(loss='binary_crossentropy',
              optimizer= opt,
              metrics=['accuracy'])

model.fit(X_train, y_train,epochs=100, batch_size=1, verbose=1)
```

```
Epoch 88/100
41994/41994 [==============================] - 168s 4ms/step - loss: 0.1816 - accuracy: 0.9367
Epoch 89/100
41994/41994 [==============================] - 170s 4ms/step - loss: 0.2345 - accuracy: 0.9362
Epoch 90/100
41994/41994 [==============================] - 168s 4ms/step - loss: 0.1873 - accuracy: 0.9367
Epoch 91/100
41994/41994 [==============================] - 168s 4ms/step - loss: 0.2435 - accuracy: 0.9367
Epoch 92/100
41994/41994 [==============================] - 170s 4ms/step - loss: 0.3206 - accuracy: 0.9373
Epoch 93/100
41994/41994 [==============================] - 169s 4ms/step - loss: 0.2495 - accuracy: 0.9371
Epoch 94/100
41994/41994 [==============================] - 167s 4ms/step - loss: 0.2976 - accuracy: 0.9367
Epoch 95/100
41994/41994 [==============================] - 166s 4ms/step - loss: 0.3050 - accuracy: 0.9371
Epoch 96/100
20694/41994 [=============>................] - ETA: 1:24 - loss: 0.1743 - accuracy: 0.9389
```

# Image Based Model

- Leveraged an existing Keras application, VGG16, for large scale image classification

| Model Type | Accuracy |
| --- | --- |
| Keras (VGG16) | 90.01% |

```
Model: "vgg16"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 256, 256, 3)]     0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 256, 256, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 256, 256, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 128, 128, 64)      0
_________________________________________________________________
block2_conv1 (Conv2D)        (None, 128, 128, 128)     73856
_________________________________________________________________
block2_conv2 (Conv2D)        (None, 128, 128, 128)     147584
_________________________________________________________________
block2_pool (MaxPooling2D)   (None, 64, 64, 128)       0
_________________________________________________________________
block3_conv1 (Conv2D)        (None, 64, 64, 256)       295168
_________________________________________________________________
block3_conv2 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_conv3 (Conv2D)        (None, 64, 64, 256)       590080
_________________________________________________________________
block3_pool (MaxPooling2D)   (None, 32, 32, 256)       0
_________________________________________________________________
block4_conv1 (Conv2D)        (None, 32, 32, 512)       1180160
_________________________________________________________________
block4_conv2 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_conv3 (Conv2D)        (None, 32, 32, 512)       2359808
_________________________________________________________________
block4_pool (MaxPooling2D)   (None, 16, 16, 512)       0
_________________________________________________________________
block5_conv1 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv2 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 16, 16, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 8, 8, 512)         0
=================================================================
Total params: 14,714,688
Trainable params: 14,714,688
Non-trainable params: 0
```

# BAYESIAN STATISTICS

- Algorithm: Bayesian logistical regression, using PYMC3

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Mathematical connection : the likelihood is the product of n Bernoulli trials,

$$\prod_{i=1}^{n} p_i^y (1-p_i)\ (1-p_i)^{1-y_i}\text{ , where } p_i = \frac{1}{1+e^{-z_i}}$$

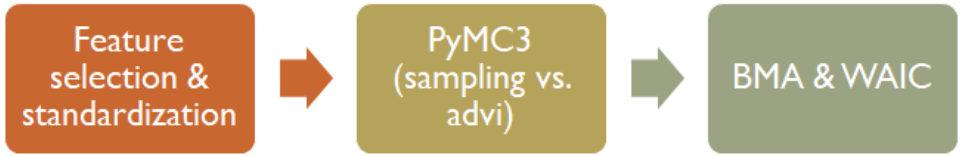
$$y_i = \beta_0 + \beta_1(\text{numPages})_i + \beta_2(\text{height})_i + \beta_3(\text{width})_i + \beta_4(\text{dim})_i + \beta_5(\text{structure})_i + \beta_6(\text{content})_i + \beta_7(\text{association})_i + \beta_8(\text{language})_i + \beta_9(\text{numChar})_i$$

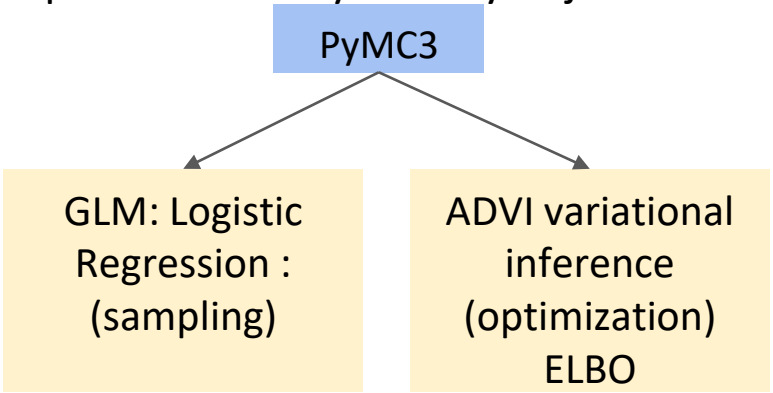
Where $y_i$ = 1 if researchPublication and $y_i$ = 0 otherwise

- Priors : default $p(\theta)$ = N(0, $10^{12}$I)

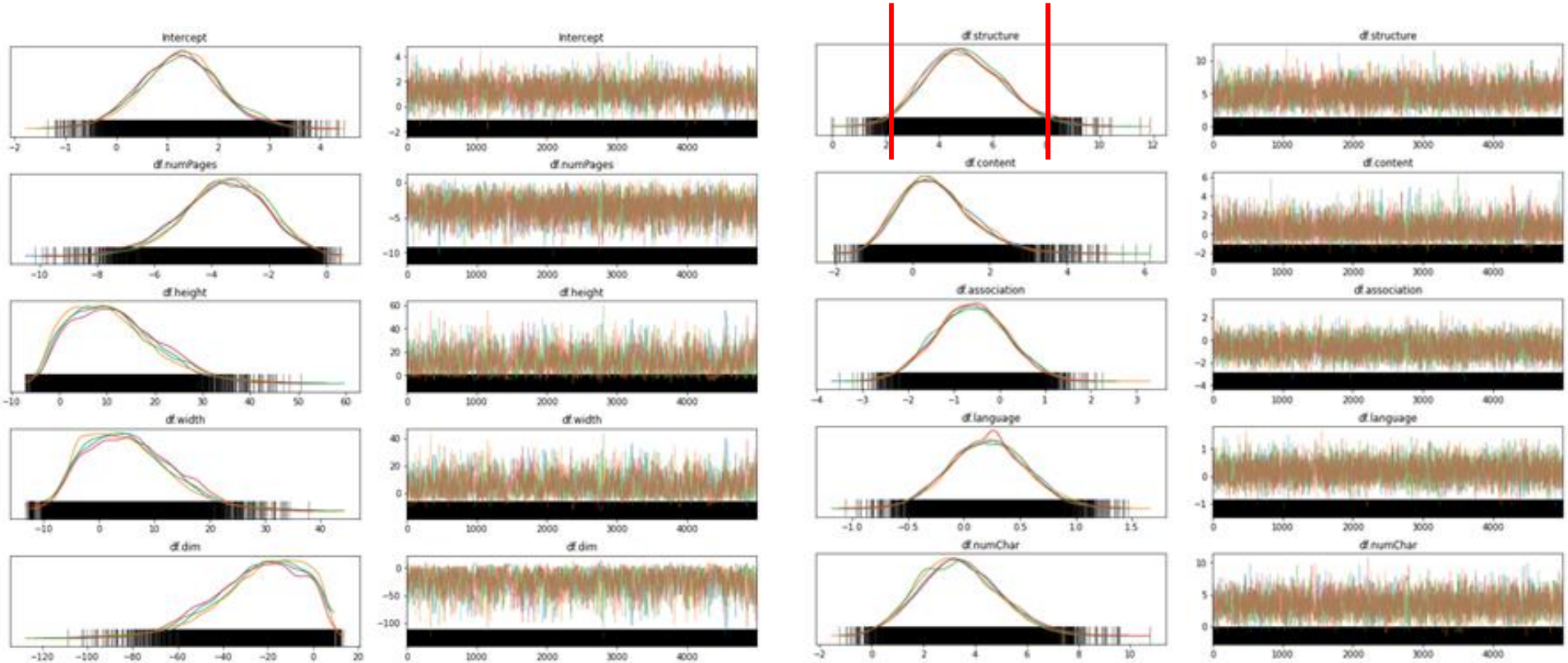
# Methods

- Total set of features considered: number of pages, height, width, dimensions of page, structure, content, association, language, number of characters
- How likely is it a **research publication** based on the selective features?
- Model comparison approach – compared different sets of features and accompanying accuracy. Given parameters for the capstone project, including speed, prioritizing a balance of the smallest number of features with acceptable accuracy is a key objective



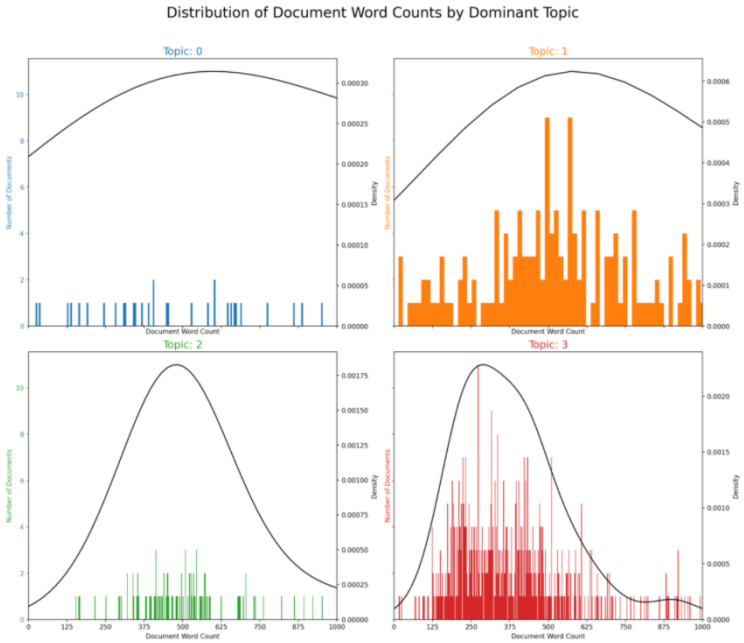



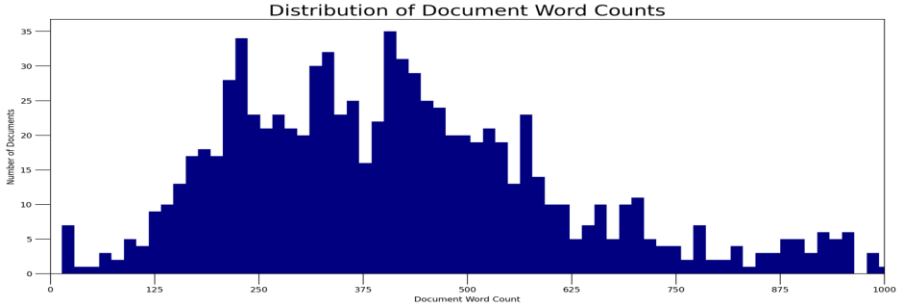
# RESULTS - Full Model (Sampling)



# Topic Modeling

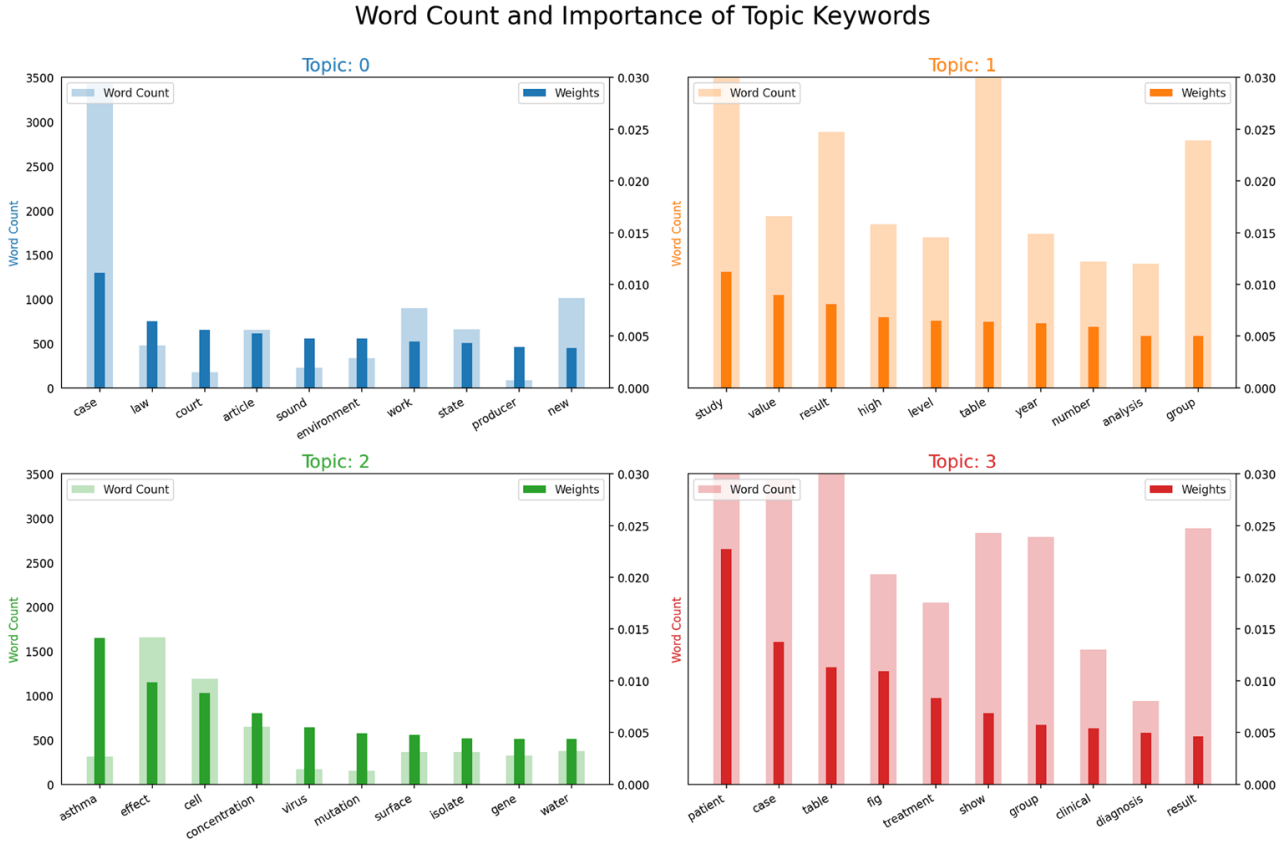
- Topic model based on LDA (Latent Dirchlet Allocation)
  - Tokenize Sentences and Clean
  - Build the Bigram, Trigram Models and Lemmatize
  - Build the Topic Model
  - Dominant topic and percentage contribution in each document
  - Frequency distribution of word counts in documents



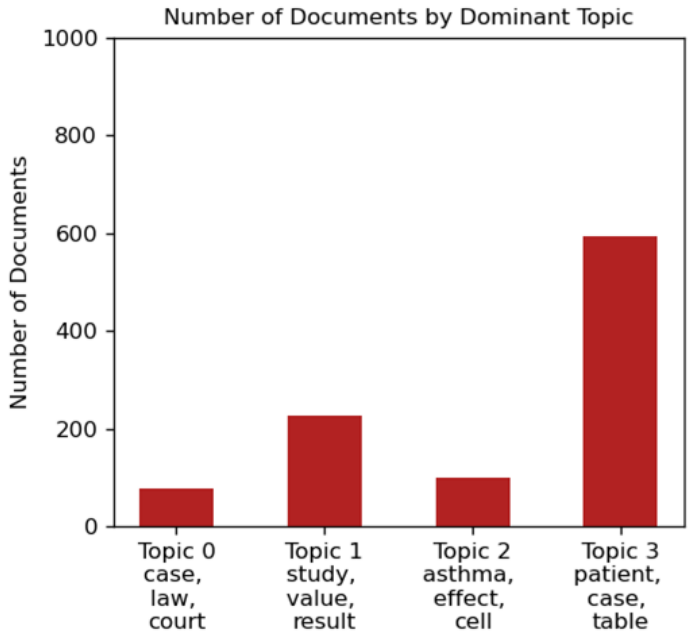



# Word Count and Weighted Keywords

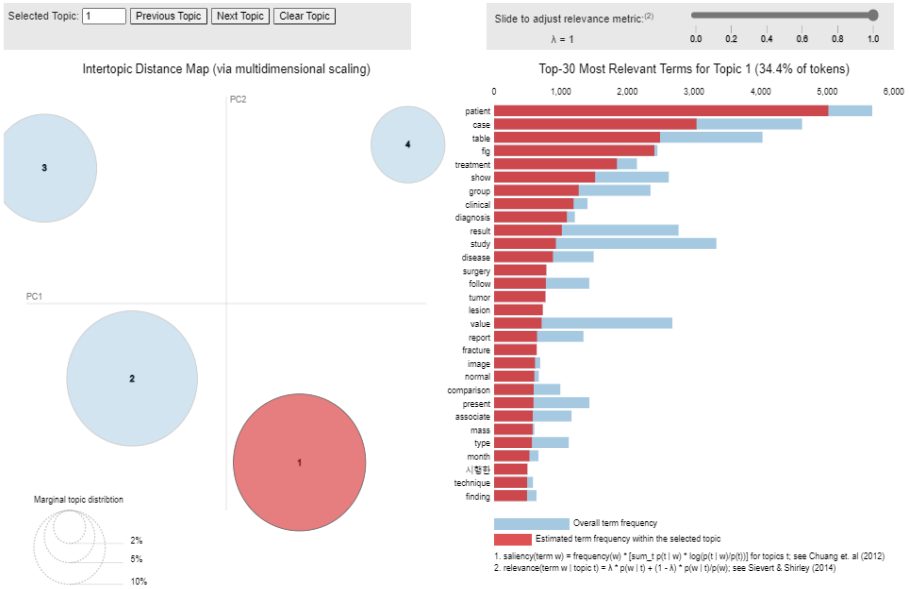
Frequency of Keywords in each Topic



# Top Topics




Assigned topic with most weight in the document



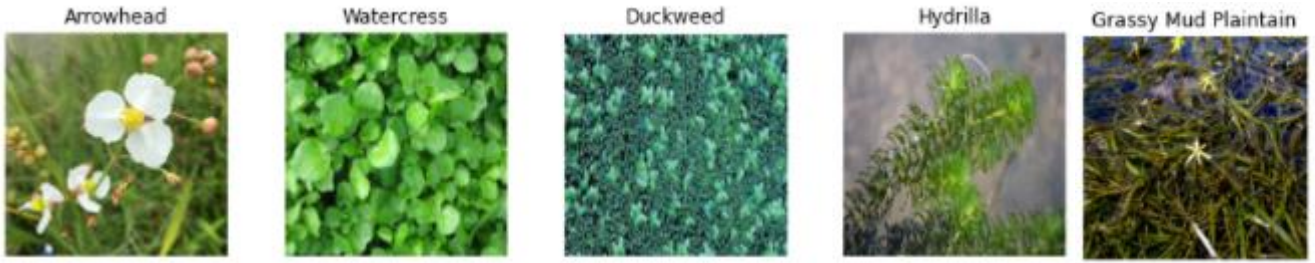

pyLDAVis Visualization of Topic Modeling

# Conclusions

- Top feature - structure
- Additional key features - number of pages, width
- Format matters, in addition to content
- Simple model may be reasonable, given comparable accuracy to others, to prioritize speed
- Compare the pros and cons of text-based vs image-based models

# Deep learning in Agriculture

- 5 Classes – one invasive and four are non-invasive plants

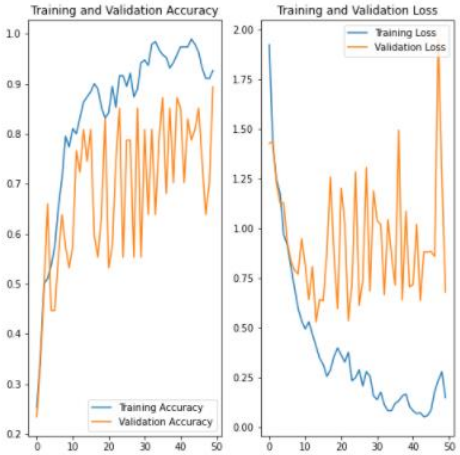


# Model 1- CNN

- Feature extractor
- Classification
- Adopt Softmax
- Dropout

```
In [39]: model.summary()
```

```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
sequential_7 (Sequential)    (None, 150, 150, 3)       0
_________________________________________________________________
rescaling_4 (Rescaling)      (None, 150, 150, 3)       0
_________________________________________________________________
conv2d_17 (Conv2D)           (None, 150, 150, 16)      448
_________________________________________________________________
max_pooling2d_15 (MaxPooling (None, 75, 75, 16)        0
_________________________________________________________________
conv2d_18 (Conv2D)           (None, 75, 75, 32)        4640
_________________________________________________________________
max_pooling2d_16 (MaxPooling (None, 37, 37, 32)        0
_________________________________________________________________
conv2d_19 (Conv2D)           (None, 37, 37, 64)        18496
_________________________________________________________________
max_pooling2d_17 (MaxPooling (None, 18, 18, 64)        0
_________________________________________________________________
dropout_4 (Dropout)          (None, 18, 18, 64)        0
_________________________________________________________________
flatten_5 (Flatten)          (None, 20736)             0
_________________________________________________________________
dense_11 (Dense)             (None, 128)               2654336
_________________________________________________________________
dense_12 (Dense)             (None, 5)                 645
=================================================================
Total params: 2,678,565
Trainable params: 2,678,565
Non-trainable params: 0
_________________________________________________________________
```



# Pretrain model

| | Train accuracy | Validation accuracy |
|---|---|---|
| Xception | 99% | 94% |
| EfficientNetB6 | 95% | 78.8% |