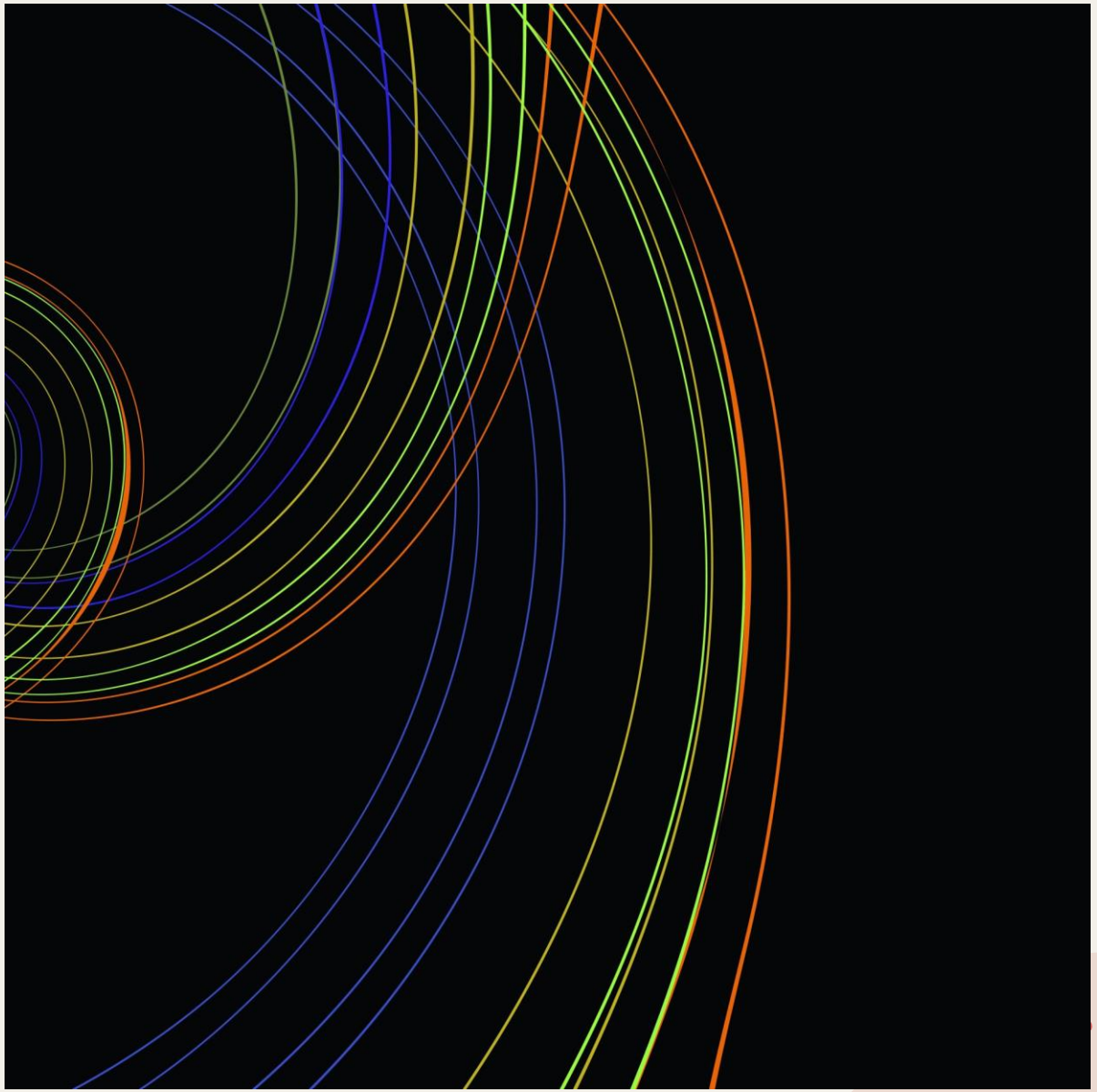
JD Industry Multi Classification



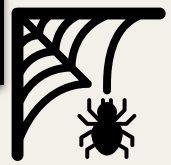
1. Objective

Techlent Search: LinkedIn for Data Scientists

- Techlent.Inc is founded to help people find data scientists' job. Techlent Search is a new service to facilitate DS job hunting.
- Candidates usually spend hours everyday browsing through all available "Data Scientist" jobs and apply for jobs that in fact do not match their expertise.
- Recruiters spend hours finding candidates and send mails to whoever titled "Data Scientist" to their roles in hand.
- We want AI to match candidates and jobs.

2. Workflow to solve this problem

Scraper 25k raw data from indeed.com



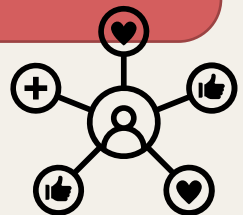
"Data Scientist" is our key word

Manual labeling to 5 categories around 1000 jd data

EDA and Feature control

3 times of Active learning each including 5 iterations:
Used 40%;15%;15% of label data

7 models (6 general ML, 1 DL) test on 30% label data
1.Logistic regression
2.Linear SVC
3.Randome Forest
4.NB
5.OVO
6.OVR
7.Deep learning lstm



Model analysis

Development to predict the reporting category of the client's resume, probablity, keywords

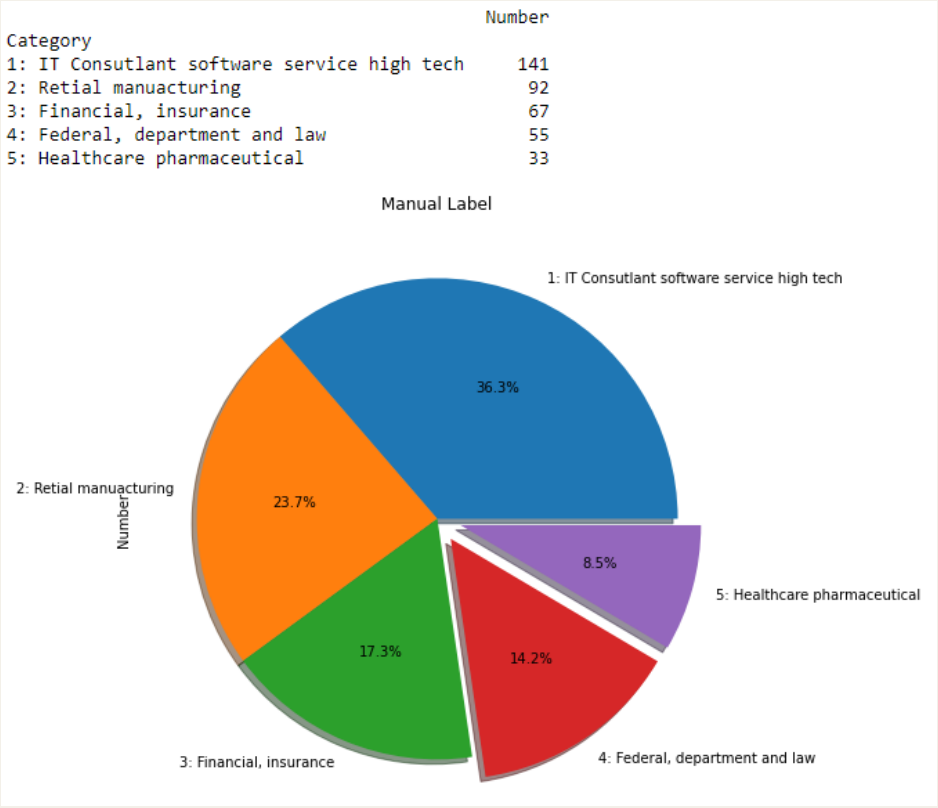
3. Data preparation and Multi Classification

Index	Category	Example
1	IT Concultant servic, software and high tech	AI platform company, IT service, Amazon, google, facebook
2	Retail & Manufacturing	Bestbuy, Walmart, verizon, AT&T, Comcast
3	Finacial, insurance	Citibank, Wells Fargo, Credit Karma
4	Federal, Depart, Law,School	National Lab, DOD, university, DOE
5	Healthcare, pharmaceutical	Unitedhealth, CVS, Hospital

- Download 10K indeed DS jobs.
- Manual label 1000 (Active learning 40%, 15%, 15% and testing 30%)

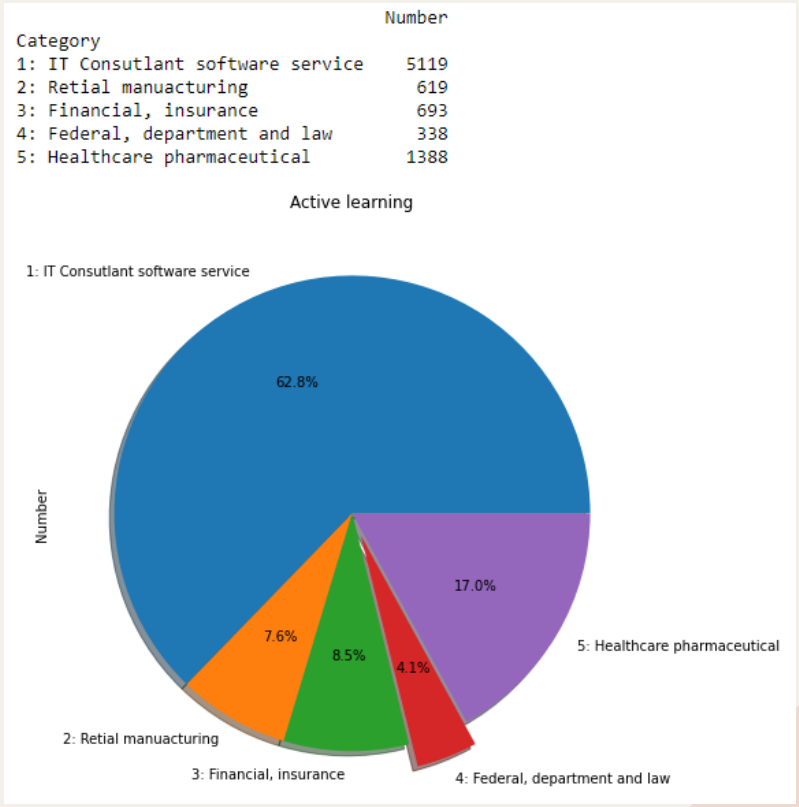
Manually Labeled (first batch 388, 3 batch total)

Category	Number
1: IT Consutlant software service high tech	141
2: Retial manuacturing	92
3: Financial, insurance	67
4: Federal, department and law	55
5: Healthcare pharmaceutical	33



After Active Learning (8157 JD)

Category	Number
1: IT Consutlant software service	5119
2: Retial manuacturing	619
3: Financial, insurance	693
4: Federal, department and law	338
5: Healthcare pharmaceutical	1388



4. General ML modeling

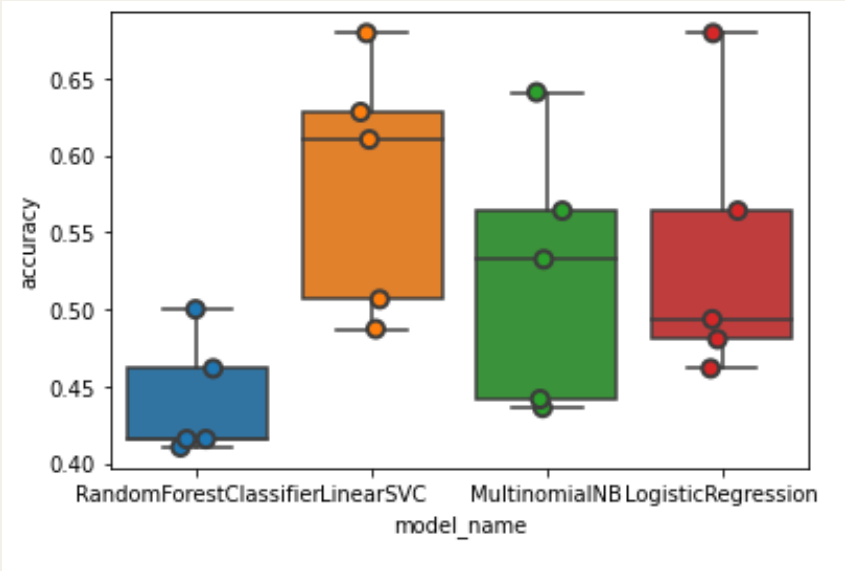
Sample category keywords

```
'Retial manuacturing':
. Most correlated unigrams:
      . seattle
      . forecasting
      . retail
      . optimization
      . customer
      . brands
      . target
      . chain
      . nordstrom
      . supply
```

```
'Healthcare pharmaceutical':
. Most correlated unigrams:
      . computational
      . drug
      . financial
      . accenture
      . biology
      . oncology
      . study
      . patient
      . patients
      . care
```

```
'Financial, insurance':
. Most correlated unigrams:
      . fraud
      . capital
      . investment
      . bank
      . lending
      . banking
      . paypal
      . credit
      . risk
      . financial
```

Base Model – before active learning



Accuracy Score:

RandomForestClassifier	0.396
LinearSVC	0.605
MultinomialNB	0.412
LogisticRegression	0.502

Six general ML models

Feature control

1. TFIDF setup, max, min_df, max_features (=1000) to avoid overfitting.
2. Deep learning, dimension, words number control
3. Logistic regression tends to be overfitting if too many dimensions, used OVR in SVC linear kernel
4. Used OVR instead of logistic regression for active learning
5. Model regularization parameters don't help much in the fine tunning process.

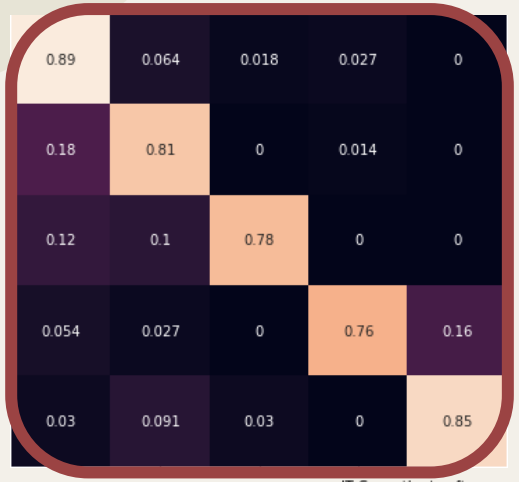
Model metrics and modeling results – after active learning

	Accuracy (Training)	Accuracy (Testing)	ROC_AUC (Ttesting)	micro_F1 (Testing)	Hamming_Loss (Testing)
MultinomialNB	0.862	0.613	0.76	0.613	0.39
Random Forest	0.680	0.675	0.63	0.404	0.60
LinearSVC	0.913	0.797	0.80	0.675	0.32
Logistic regression	0.979	0.831	0.89	0.831	0.17
One vs one (LinearSVC)	0.989	0.785	0.87	0.785	0.22
One vs rest(SVC Linear kernel)	0.985	0.801	0.88	0.801	0.20

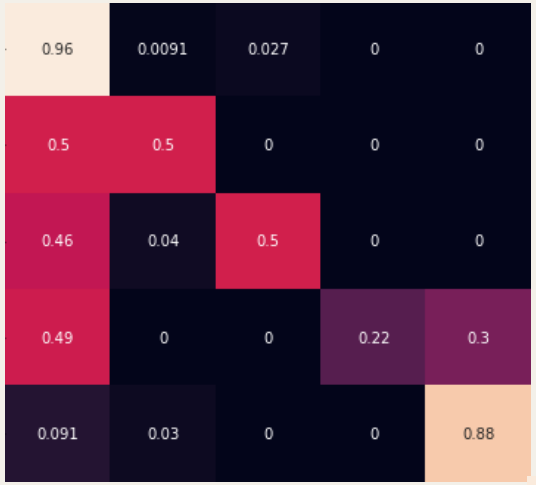
Training on about 8500 active learning labelled jd, testing on 300 labelled jd.

Confusion Matrix

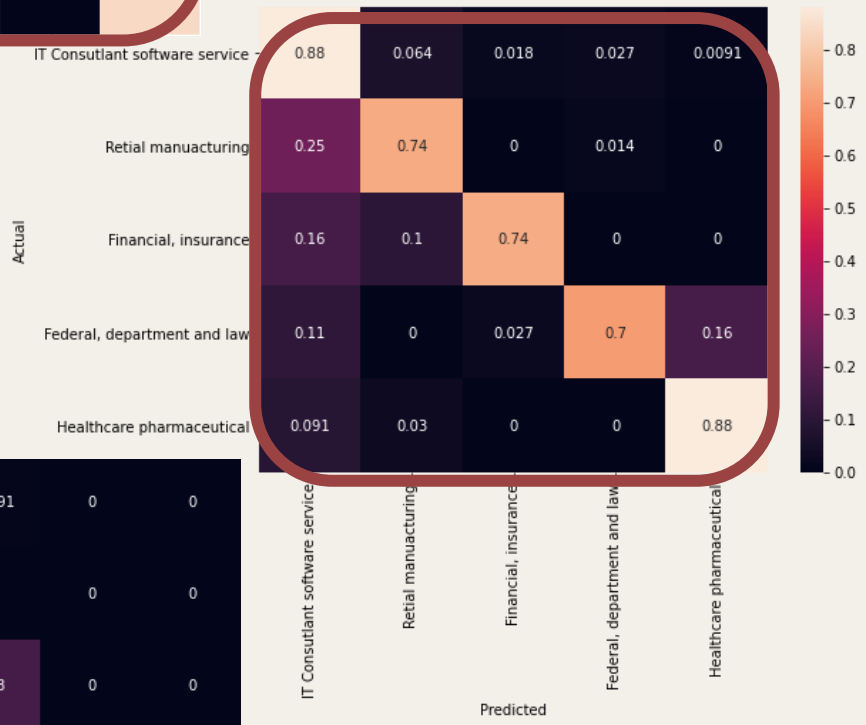
Logistic regression



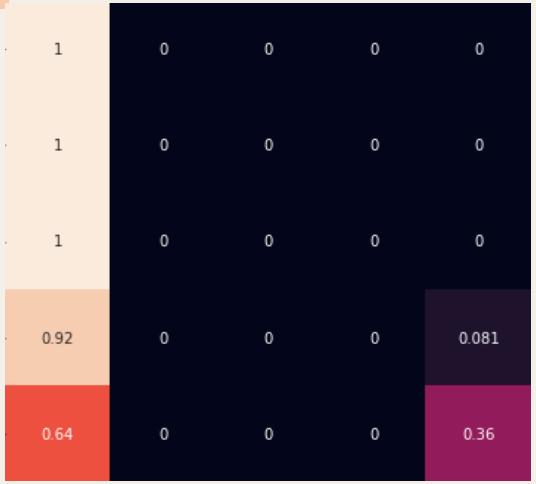
LinearSVC



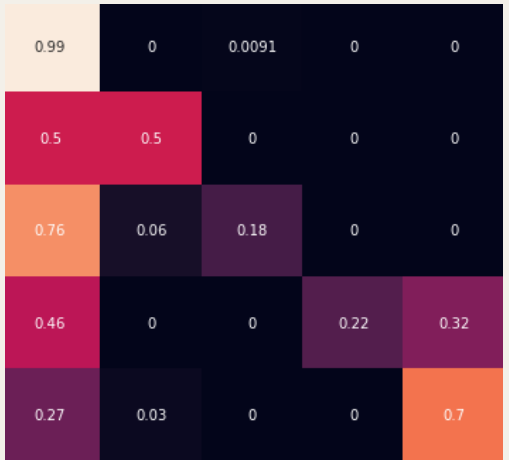
OVR+SVC linear kernel



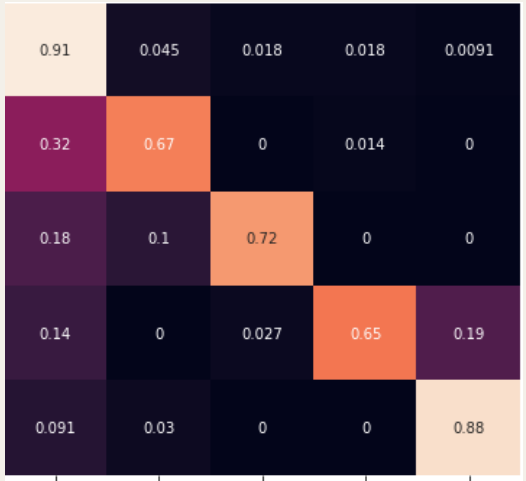
RF



NB

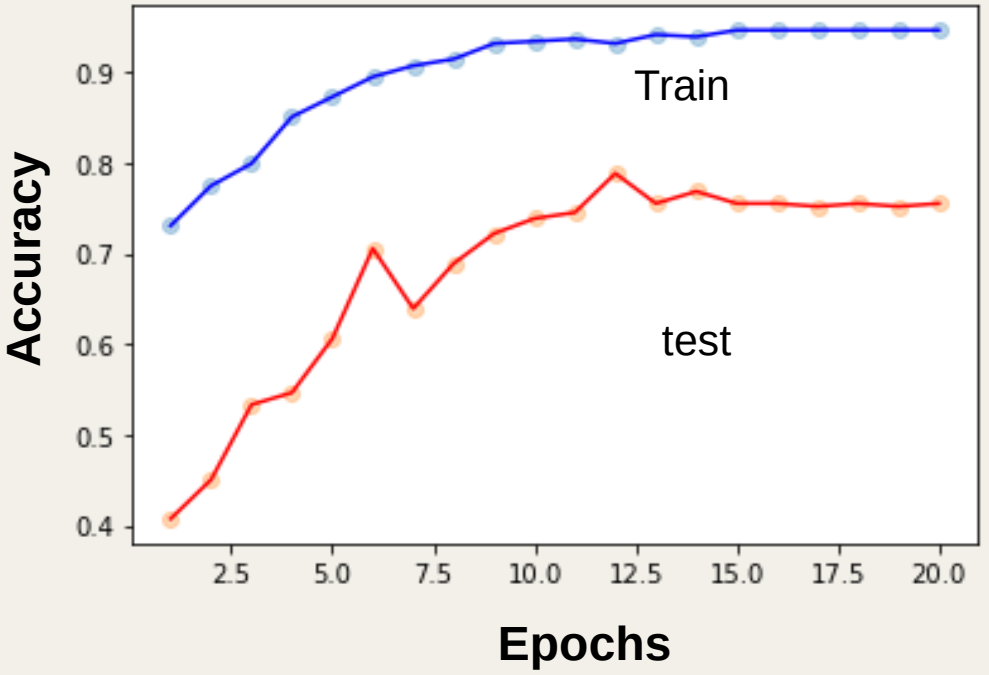


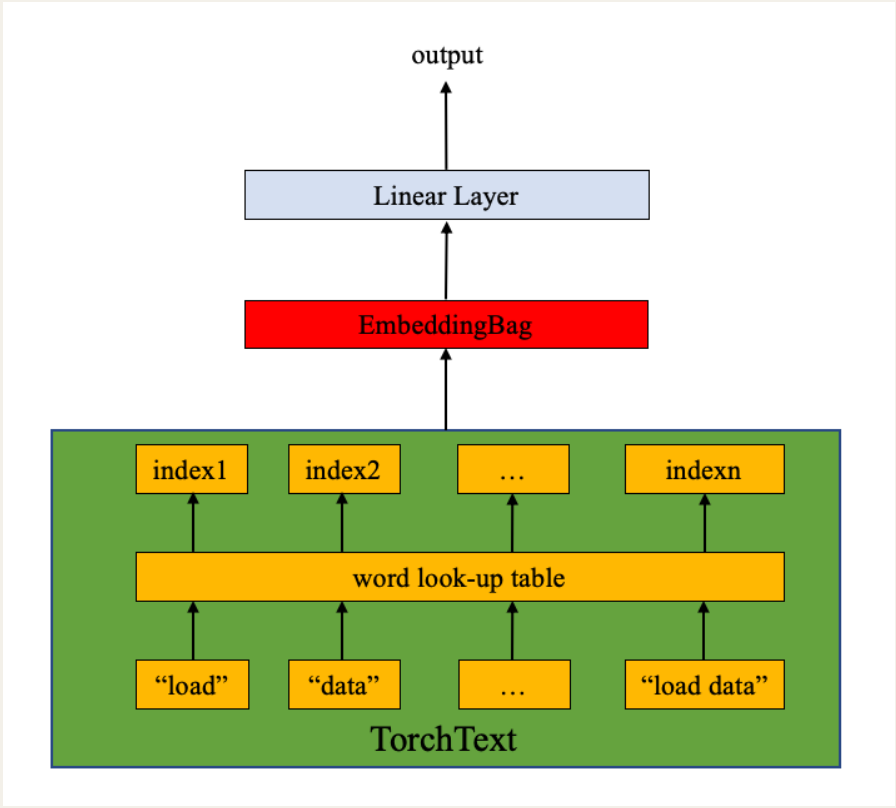
OVO+LinearSVC



5. Deep Learning LSTM (1)

1. PyTorch
2. Embedding dimension: 96
3. Training accuracy (0.94) and testing (0.755)
4. Performance in the middle above 6 ML models.





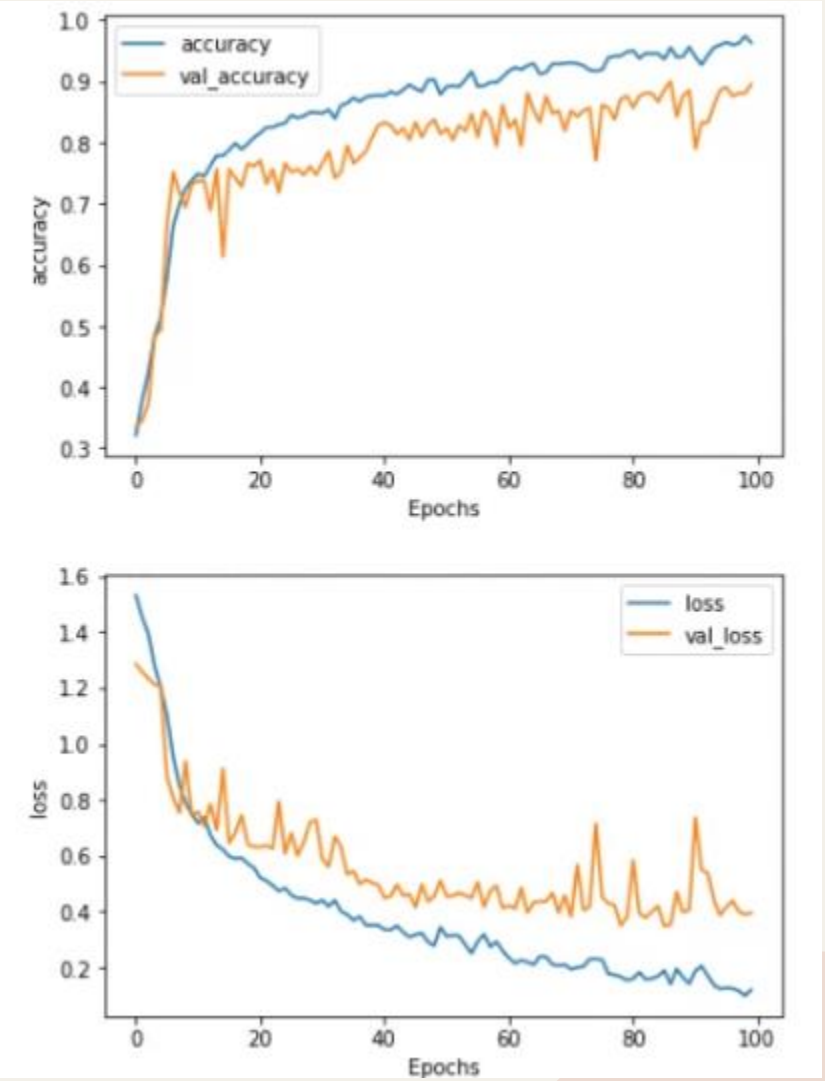
	Accuracy
1. IT Consutlant software service	92.7%
2. Retial manuacturing	62.5%
3. Financial, insurance	62.0%
4. Federal, department and law	54.1%
5. Healthcare phacvrmaceutical	90.9%

Deep Learning LSTM (2)

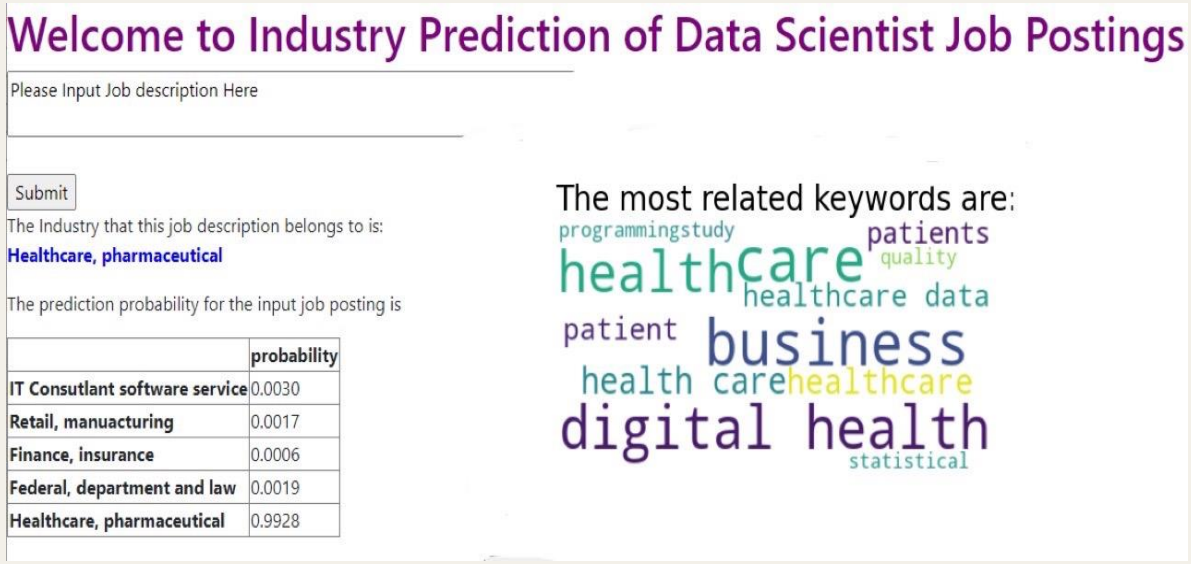
1. Keras LSTM using Glove embedding (840B.300d)
2. Max word length 150 for each JD.
3. Training accuracy (0.96) and testing (~0.6)
4. Limited data, data quality, no detailed fine tuning.

```
model = JD_classfier((maxLen,))
model.summary()
```

```
Model: "model_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_7 (InputLayer)         (None, 4000)              0
_________________________________________________________________
embedding_3 (Embedding)      (None, 4000, 300)         7142100
_________________________________________________________________
lstm_16 (LSTM)               (None, 4000, 64)          93440
_________________________________________________________________
dropout_9 (Dropout)          (None, 4000, 64)          0
_________________________________________________________________
lstm_17 (LSTM)               (None, 4000, 64)          33024
_________________________________________________________________
dropout_10 (Dropout)         (None, 4000, 64)          0
_________________________________________________________________
lstm_18 (LSTM)               (None, 64)                33024
_________________________________________________________________
dense_6 (Dense)              (None, 5)                 325
=================================================================
Total params: 7,301,913
Trainable params: 159,813
Non-trainable params: 7,142,100
_________________________________________________________________
```



6. Result - Deployment



Director, Data Science
Carrum Health
San Francisco, CA
https://www.indeed.com/rc/clk?jk=ce19e9fb31e795e4&fccid=87720e23c0f575f1&vjs=3
At Carrum, we live and breathe the mission of transforming the healthcare system to create an unmatched experience for patients. If you are passionate about changing healthcare and want to finally get rid of surprise bills, poor quality, and high prices, while thriving in an entrepreneurial, cutting-edge environment, we would love to connect with you.

In 2014 Carrum reinvented the Center of Excellence (CoE) category in digital health. We are the only company in this space with a digital platform and mobile app powering this novel marketplace. Today, most of the US population lives within a 3 hours driving distance of a Carrum CoE and our providers rank in the top 10% nationally. Our team's execution has been recognized by the venture community and we've raised more than $50M in aggregate from investors like Tiger Global Management and Wildcat Ventures. Our impact has been externally proven in a 2021 RAND Corporation study and featured as a Harvard Business School (HBS) case study .

The Director of Data Science role will report to the Chief Product Officer and be responsible for leading a team of data analysts and scientists. The ideal candidate is someone who builds teams, executes company-wide data strategy, and loves the intensity of a high-growth startup.

You're excited about this opportunity because you will...

Develop, propose, and execute on data science initiatives and roadmap. Define best practices and innovations in the personalization, predictive modeling, and machine learning space.
Build a stellar data team. Help them grow through effective mentorship and lead them to deliver on the roadmap.
Spearhead the development of a robust data foundation and infrastructure. Help implement data architecture, processes, systems and self-service tools for the company.
Partner and influence functional leadership to ensure team priorities and output directly impacts compan

7. Summary

- *High-quality data* are the precondition for analyzing and modeling.
- Feature selection and quantity control are crucial for applying NLP methods.
- Logistic regression, OVR, and OVA have better performance. Therefore, We would recommend using one of these.
- Model regularization parameter tunning didn't help that much in this project.

Next Steps

1. Consider combining job classification with industry classification together to make a more comprehensive recommender system.
2. Connect the locations with different categories of jobs to help candidates target jobs in potential areas.

Q&A