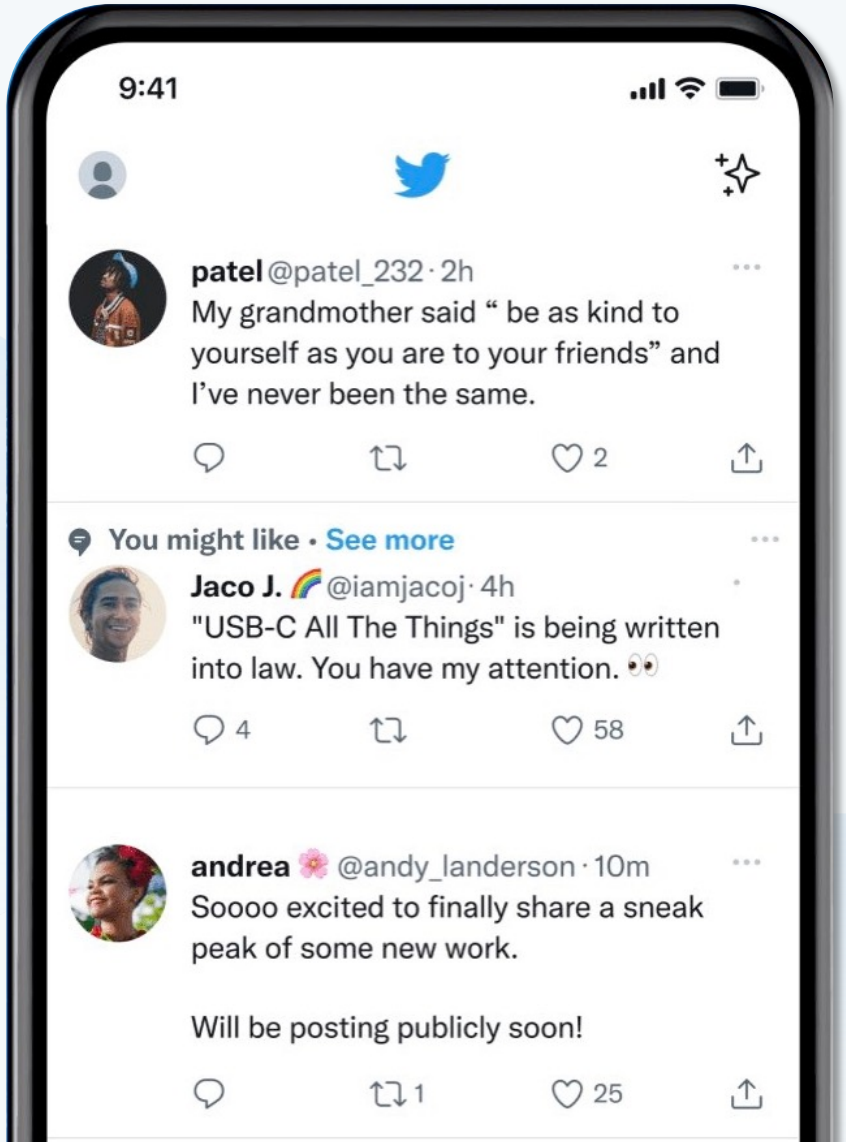
Hashtag Recommender

On Twitter

#TagFinder Group

Huaiyue Chang	6365696861	Manxin Wang	9537808576
Qianran Ma	7798937244	Wanning Li	5898212888
Yanan Zhou	4366462176	Yitong Wang	1299603893



Content

01 Problem Statement

Presenter: Qianran Ma

What we will do?

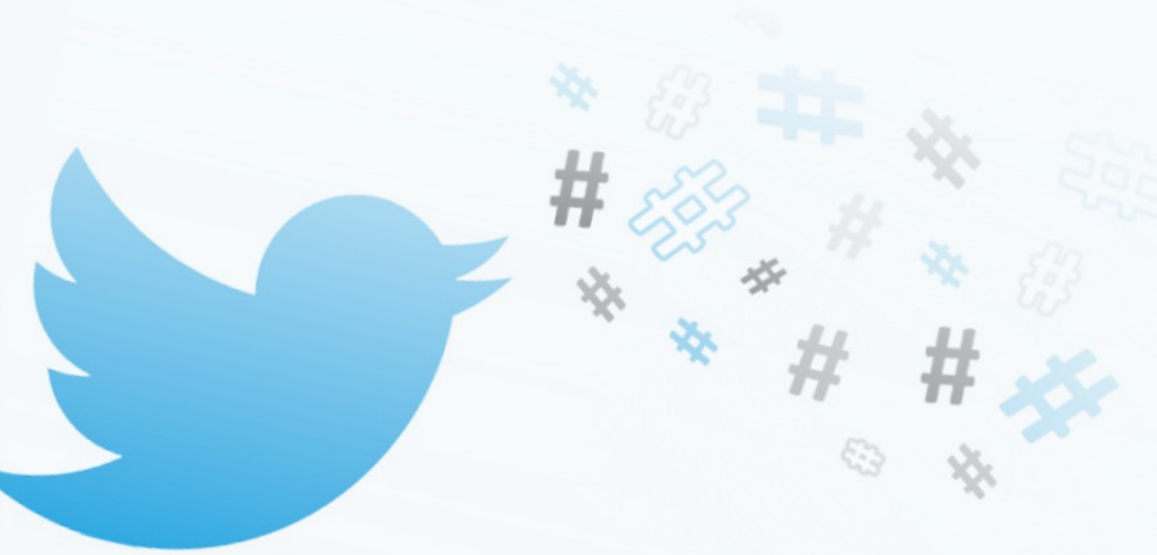
Analyze tweets text and recommend top possible hashtags.

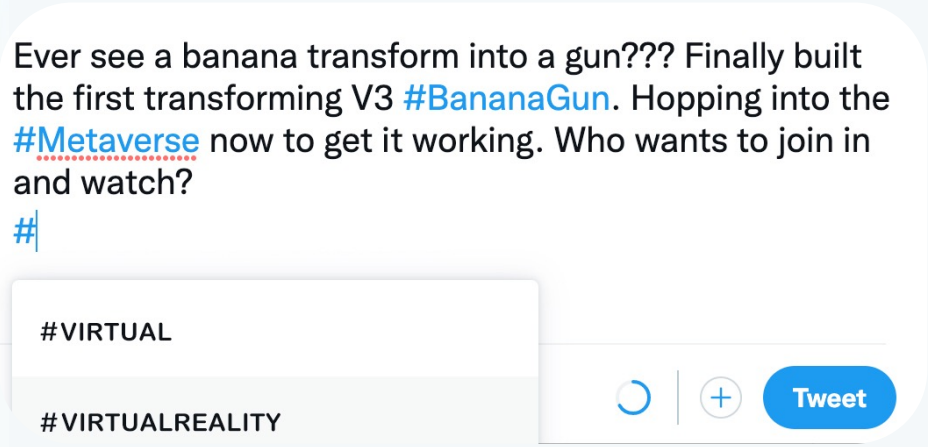
Why we are doing it?

Using hashtags in tweets can greatly increase user engagement, establish brand images and attract target users.

How to acchieve it technically?

Build a set of hashtags and treat it as a multilabel classification problem.





02 Data Preparation – Data Collection

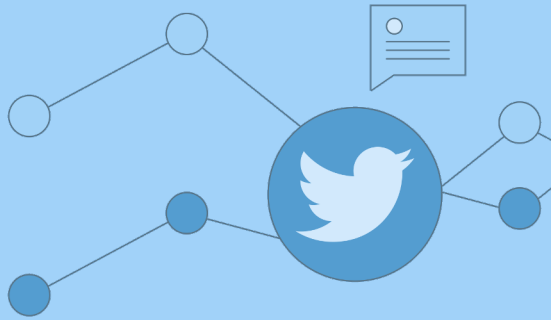
Presenter: Qianran Ma

Source	Twitter
Format	DataFrame saved as csv file with 4 columns [text, created time, hashtags, entity name]
Volume	More than 700k collected

	text	created_time	hashtags	entity_name
0	19/ Also Wagner PC r now recruiting the worst...	2022-10-28 19:56:50+00:00	[SBU]	(RED)
1	#SundarC #khushsundar @Udhayanidhi Stalin \n#A...	2022-10-28 18:01:30+00:00	[SundarC, khushsundar, AvniCinemax, BenzzMedia...	(RED)
2	Highlights from @Red Crane Studios PH @PBAmoto...	2022-10-28 11:56:29+00:00	[pinoyvlogger, youtuber, youtube, pinoy, vlogg...	(RED)
3	Thank you @P_MCDOWELL for supporting this year...	2022-10-28 11:16:21+00:00	[MoreThanARun, RedRunLdn, TakeActionLaceUp, Mo...	(RED)
4	On 26/11/22 the World AIDS Day Red Run Returns...	2022-10-28 11:01:17+00:00	[redrunuk, worldaidsday, takeactionlaceup, mor...	(RED)
...	...	...	...	...
12	bump! still avail, just hit me up 😻 #pasarseve...	2022-10-28 06:11:09+00:00	[pasarseventeenmy, pasarseventeen]	letgo
13	WTS Ungu 26th Anniversary Concert Live in Mala...	2022-10-28 04:51:46+00:00	[ungu, unguliveinKL, ticketconcert, WTS]	letgo
14	wtb unsealed red velvet albums without photoca...	2022-10-28 04:49:29+00:00	[pasarredvelvet, pasarrv, pasarluvies]	letgo
15	PLAYER_reset_kick_buffer\nPLAYER_letgo_ball\nP...	2022-10-28 04:18:24+00:00	[fifa99]	letgo
16	Bump😊\nHj starriver ❌\nChan lotte still availa...	2022-10-28 03:47:54+00:00	[pasarskz, pasarstraykids]	letgo

Details

- Link to data: raw data
- Specific domain: Brand
- Query filters:
  - Domain & Entity pairs: provided by official csv, non English entities filtered, 2378 pairs left
  - Is not retweet, has hashtag, is English
- Recent search function: data within 7 days available
- Data range: 10/25/2022-11/05/2022



02 Data Preparation – Data Preprocessing

Presenter: Qianran Ma

Step 1: Data Cleaning

Text data cleaning:

- Remove url, html, emoji
- Lower letters
- Remove punctuations except ",.!?:;"
- Word lemmatization
- Remove hashtags at the end of the text, but retain hashtags in the middle of text

Hashtag cleaning:

- Remove stopwords
- Lower letters
- Remove words not composing of all letters

Step 2: Data Selection

Select the data whose hashtags are in the top 50 hashtag set.

Step 3: Encoding

Hashtag encoding:

- Multi-hot encoding

Text data encoding:

(Contained in the following model training part)

Step 4: Dataset Splitting

Total data: 137k

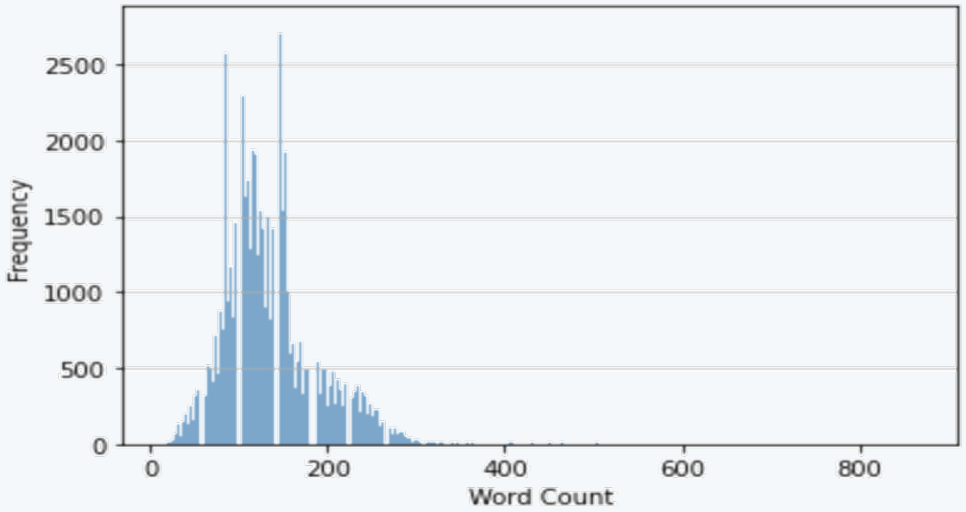
- Training dataset (50%): **68,664**
- Validation dataset (30%): **40,960**
- Test dataset (20%): **27,703**

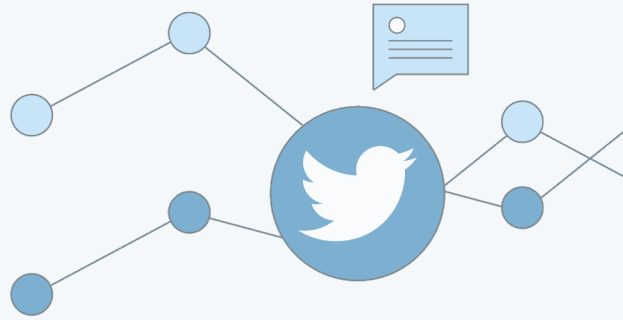
Data Preparation – Data Description

Presenter: Qianran Ma

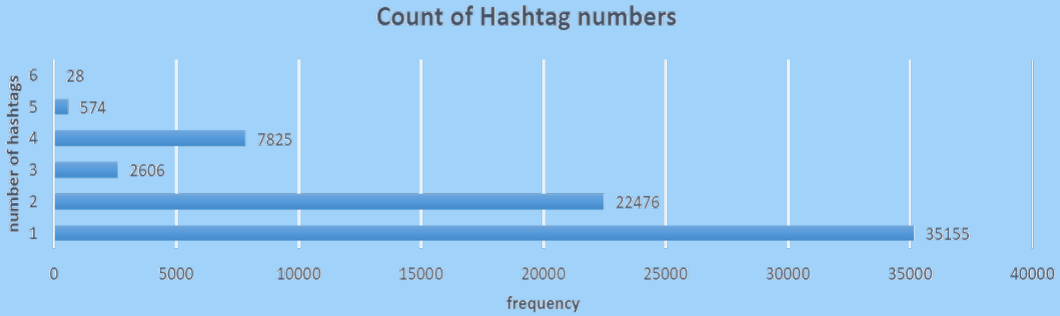
X_train: Tweets (Text Data)

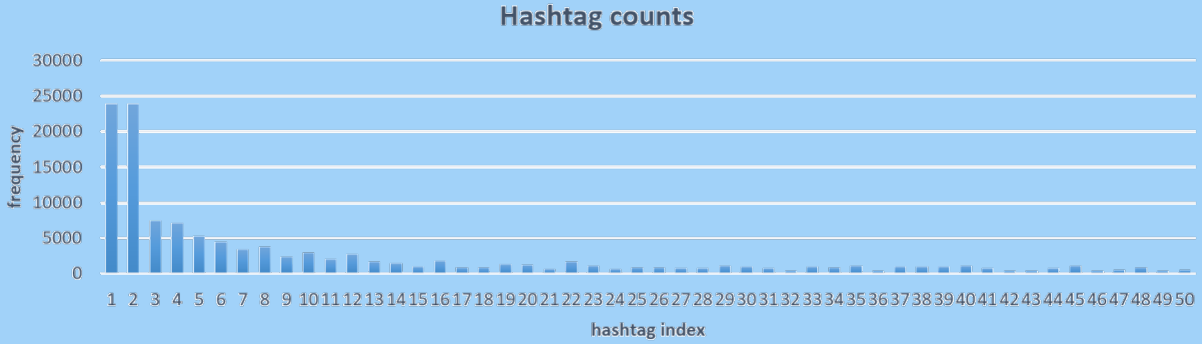
count	68664.000000
mean	134.212455
std	52.587363
min	12.000000
25%	100.000000
50%	125.000000
75%	154.000000
max	869.000000





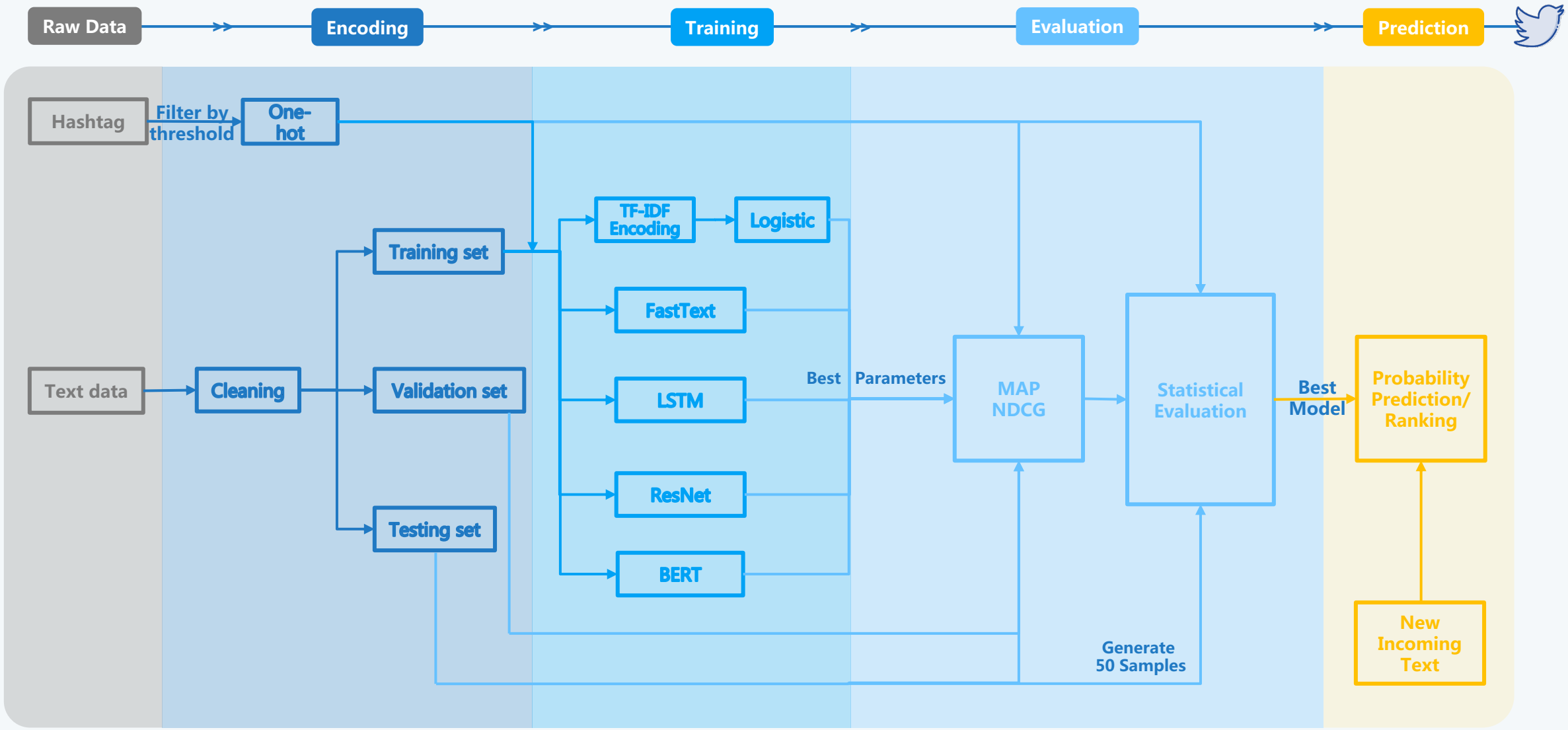
y_train: Hashtags (Multi-Hot Encoding)





03 Models – Pipeline Structure

Presenter: Qianran Ma



03 Models – Random Baseline

Presenter: Qianran Ma

Multilabel classification problem

- Regard each label as an independent Bernoulli Distribution
- Using the frequency of each label in training set as the estimated probability
- Rank the labels by the estimated probability, which means the prediction of this model for all the tweets are the same
- Calculate NDCG@3/5/10 and MAP@3/5/10 as baseline metrics

Predicted probabilities by random model

	0	1	2	3	4	5	6	7	8	9	...	40	41	42	43	44	45	46
0	0.348101	0.347955	0.10815	0.10263	0.075542	0.064444	0.049837	0.054512	0.034967	0.043487	...	0.009976	0.006437	0.005578	0.011127	0.016049	0.005024	0.007238
1	0.348101	0.347955	0.10815	0.10263	0.075542	0.064444	0.049837	0.054512	0.034967	0.043487	...	0.009976	0.006437	0.005578	0.011127	0.016049	0.005024	0.007238
2	0.348101	0.347955	0.10815	0.10263	0.075542	0.064444	0.049837	0.054512	0.034967	0.043487	...	0.009976	0.006437	0.005578	0.011127	0.016049	0.005024	0.007238
3	0.348101	0.347955	0.10815	0.10263	0.075542	0.064444	0.049837	0.054512	0.034967	0.043487	...	0.009976	0.006437	0.005578	0.011127	0.016049	0.005024	0.007238
4	0.348101	0.347955	0.10815	0.10263	0.075542	0.064444	0.049837	0.054512	0.034967	0.043487	...	0.009976	0.006437	0.005578	0.011127	0.016049	0.005024	0.007238
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

NDCG@3/5/10 and MAP@3/5/10 of random model:

	NDCG@3	NDCG@5	NDCG@10	MAP@3	MAP@5	MAP@10
Validation	0.3544	0.3825	0.4433	**0.9851**	0.8533	0.6566
Test	0.3516	0.3799	0.4403	**0.9850**	0.8518	0.6556

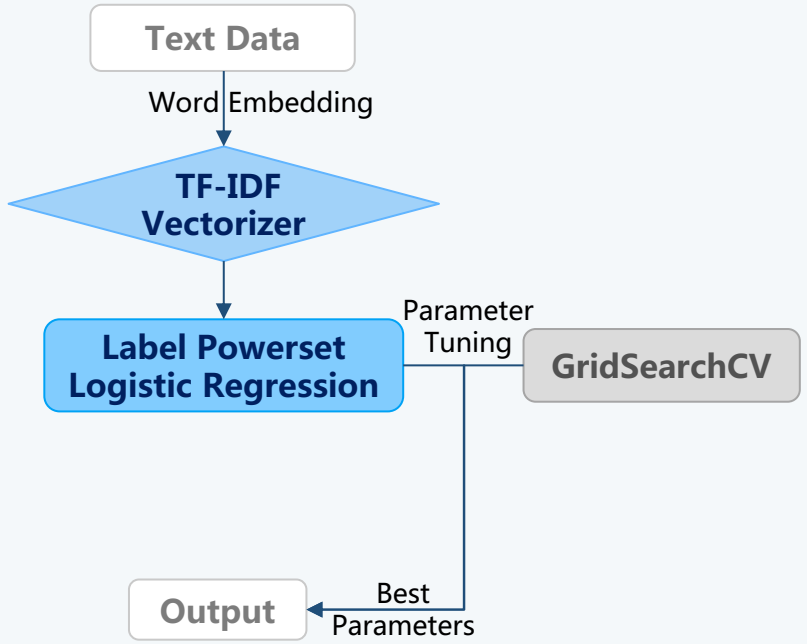
Label	Predicted probability
0	0.3481009
1	0.34795526
2	0.10814983
3	0.1026302
4	0.07554177
5	0.06444425
6	0.04983689
7	0.05451183
8	0.03496738
9	0.04348713
10	0.02908365
11	0.03927822
12	0.0241757
13	0.02022894
14	0.01457824
15	0.02590877
16	0.0124665
17	0.01175288
18	0.01856868
19	0.01766573
20	0.00895666
21	0.02413201
22	0.01504427
23	0.00902948
24	0.01175288

Label	Predicted probability
25	0.01165094
26	0.01099557
27	0.01019457
28	0.01526273
29	0.01328207
30	0.01128685
31	0.00630607
32	0.01334032
33	0.01160725
34	0.01615111
35	0.00632063
36	0.01438891
37	0.01338401
38	0.01341314
39	0.01533555
40	0.00997612
41	0.00643714
42	0.00557789
43	0.01112665
44	0.01604917
45	0.00502447
46	0.00723815
47	0.01168007
48	0.00535943
49	0.00696144

03 Models – Tf-idf + Logistic Regression

Presenter: Yiatong Wang

Pipeline



Model Settings

Tf-idf: `{"max_len": 500}`

Label Powerset:

Multilabel Classification → Multiclass Classification

Logistic Regression:

```
Loss Function = Cross Entropy
params = {
"penalty": ['none', 'l2'],
"C": [0.1, 1, 10],
"max_iter": [50, 100]
}
```

GridSearchCV: `{"scoring": f1, "cv": 2}`

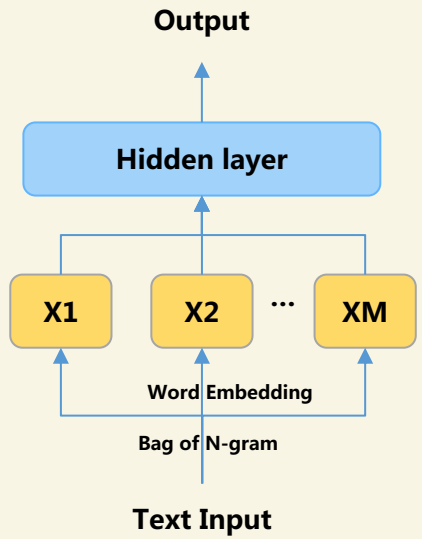
Results

	NDCG@3	NDCG@5	NDCG@10	MAP@3	MAP@5	MAP@10
Validation	0.9019	0.9105	0.9174	0.9687	0.9508	0.9310
Test	0.9038	0.9109	0.9173	0.9681	0.9529	0.9338

03 Models – FastText

Presenter: Huaiyue Chang

FastText supervised model structure



Auto validation (F1)
Average by dimension

Time and Space Efficienct

- Bag of N-gram
- Hierarchical softmax based on Huffman coding tree (O(logN)) rather than softmax (O(N))

Hidden layer → Multinomial Logistic Regression

Validation

Autotune randomly update set of parameters in a fixed range and validate model based on f1 score.

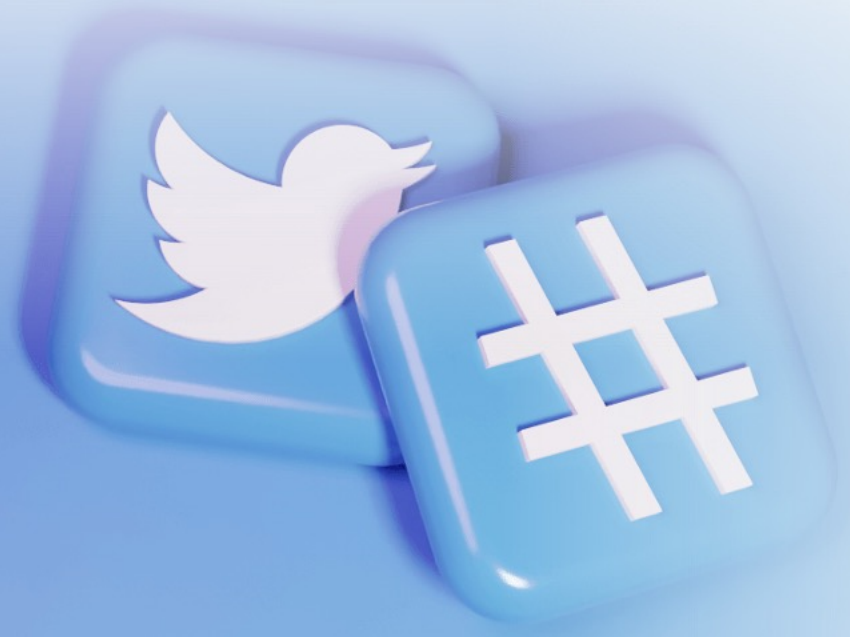
Results

	NDCG@3	NDCG@5	NDCG@10	MAP@3	MAP@5	MAP@10	Time(s)*	Size(MB)
50 labels	0.8337	0.7379	0.6591	0.8587	0.6814	0.5659	359	17MB
* validation time included. **50-label model**: epoch = 47, dimension = 55, N-grams = 1								

03 Models – Deep Learning Models (LSTM, ResNet, BERT)

Presenter: Yanan Zhou

Difficulties — No resources and time to tune parameters and add more layers.



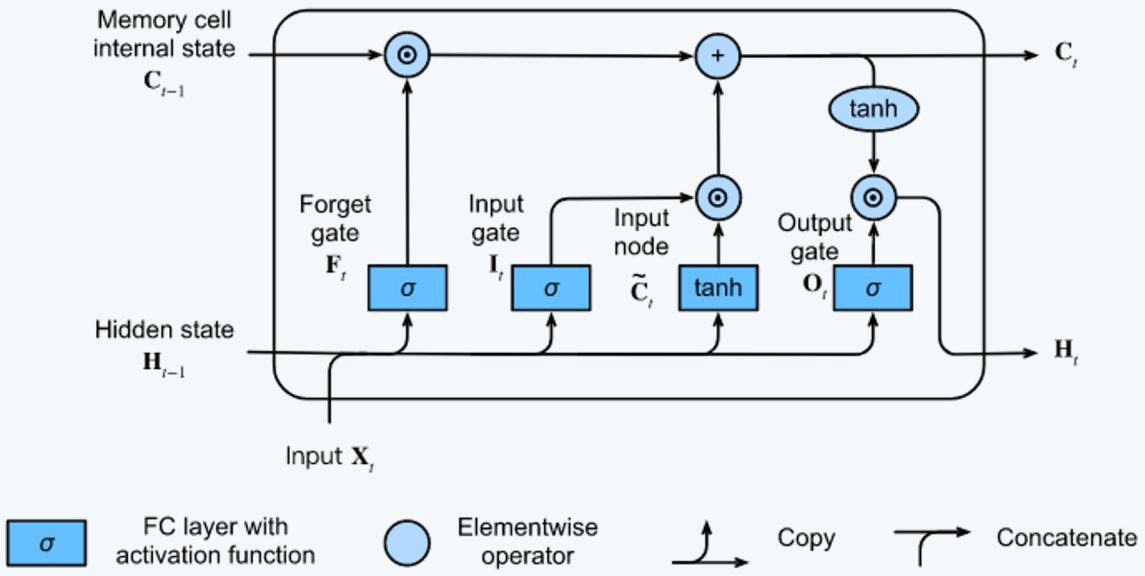
Common Settings

Sequence Length = 200	Truncate the embedding sequence if longer, pad with 0 if shorter
Batch Size = 1024	Drop last batch if less than 1024 for training and keep for validation and test
Epoch = 30	Finish training and store model checkpoint for each 10 epochs
IDF Weight	For binary cross entropy loss function to deal with imbalanced data
Threshold = 0.3	1 if the predicted probability >= 0.3, else 0

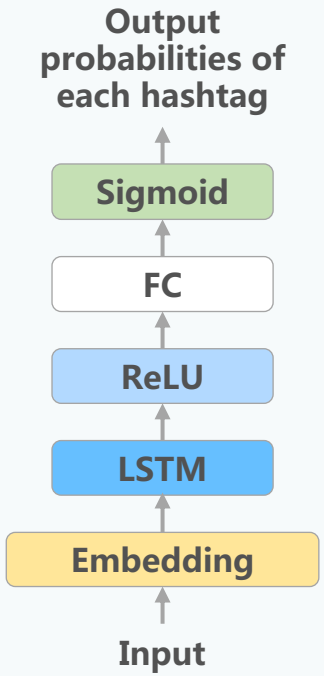
03 Models – LSTM(Long Short-Term Memory)

Presenter: Yanan Zhou

The Structure of a LSTM Unit



The Structure of LSTM Model



LSTM

Embedding
Input size = 200
Output size = 50

LSTM
Hidden size = 50,
Bidirectional = False
Number layer = 1

FC
Input size = 50
Output size = 50

Sigmoid

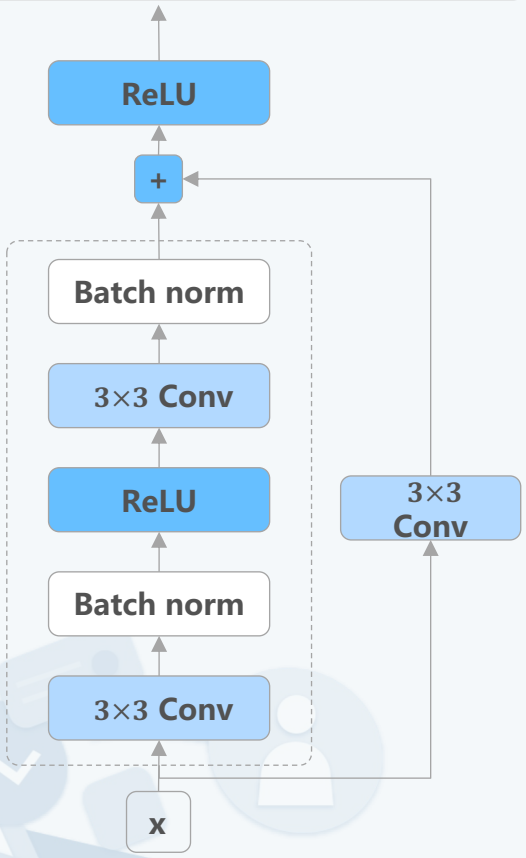
Optimization

Xavier weight initialization
Adam

03 Models – Text ResNet (Residual Networks)

Presenter: Yanan Zhou

The Structure of a Residual Block



The Structure of Text ResNet

Residual Block:

3×3 Conv

Conv1d, Padding = 1, Stride = 1

Text Resnet:

Constant Embedding

GloVe

7×7 Conv

Conv1d, Padding = 3, Stride = 2

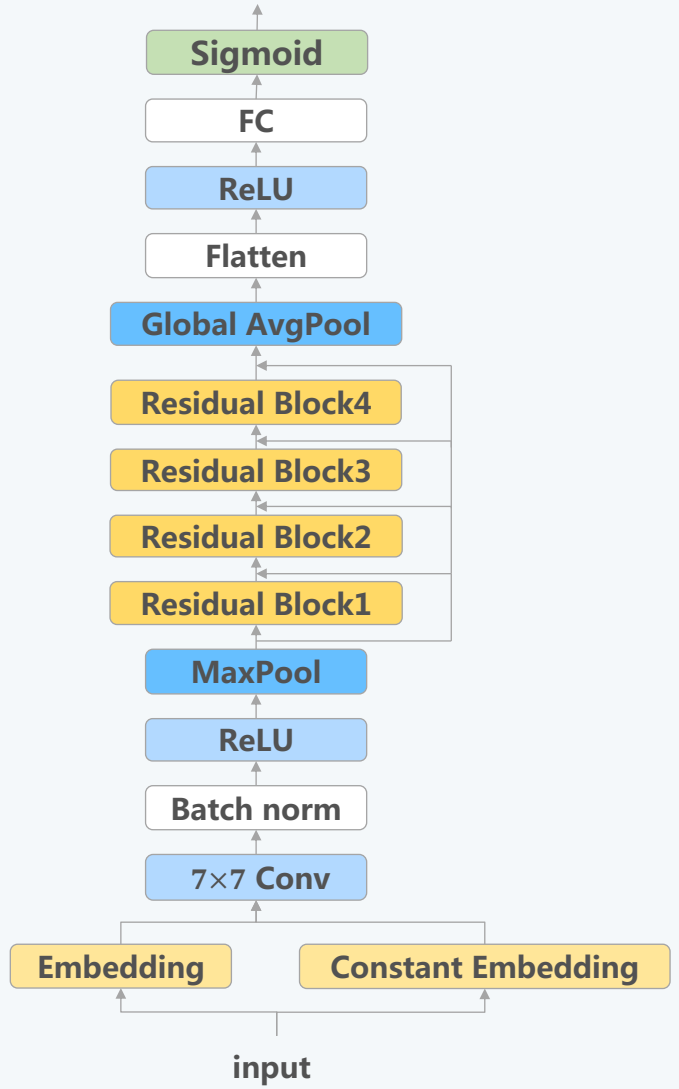
MaxPool

1d, Kernel size = 3, Padding = 1, Stride = 2

Optimization:

- Dropout = 0.5
- Xavier weight initialization
- Batch normalization
- Adam

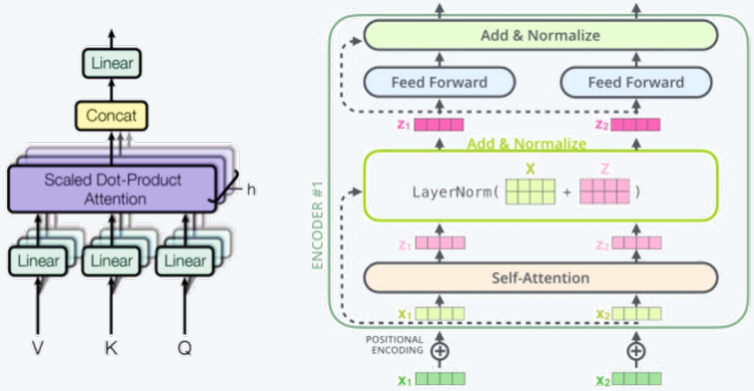
Output probabilities of each hashtag



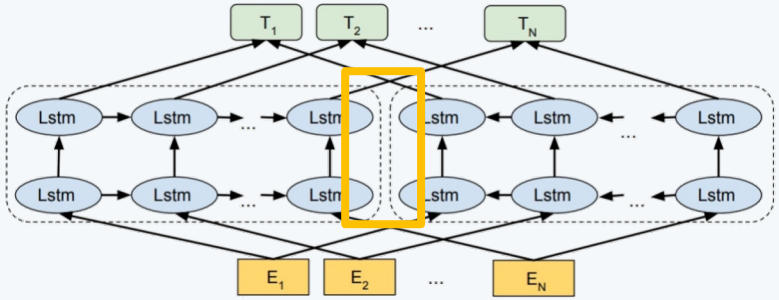
03 Models – BERT (Bidirectional Encoder Representations from Transformers)

Presenter: Yanan Zhou

Transformer Encoder

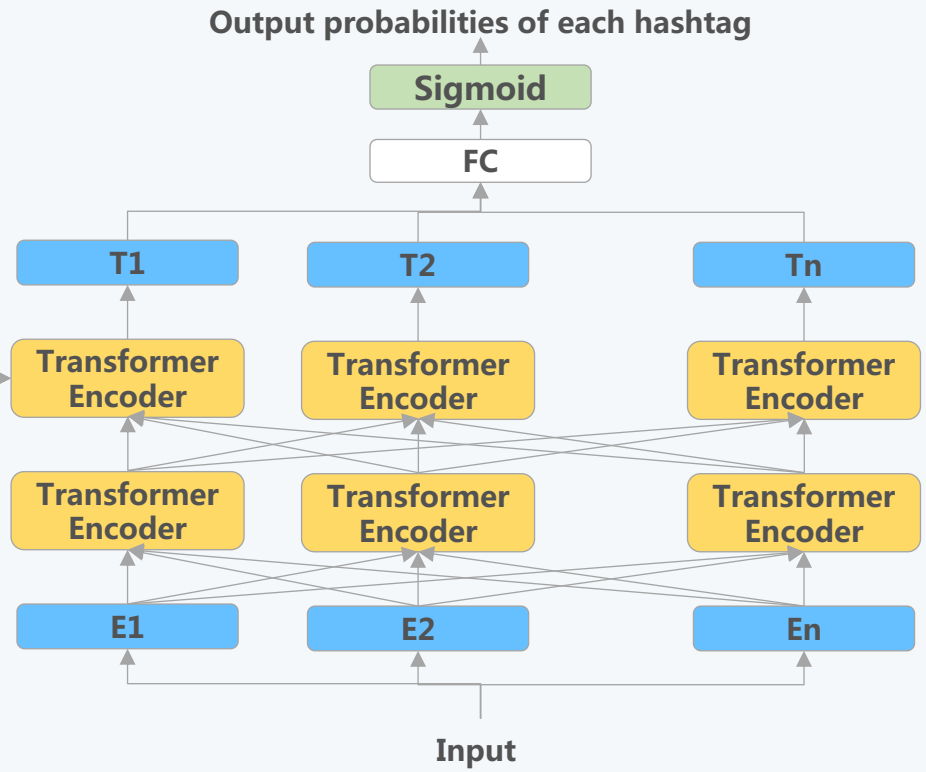


EMLo



Optimize the bi-direction learning process

BERT Structure



Pre-trained BERT

Number layer: 12

Number heads (self-attention): 12

Hidden size: 768

Total parameters: 110M

New Added Layer

Input size: 768

Output size: 50

03 Models - Performance of Deep Learning Models

Presenter: Yanan Zhou

- Save model every 10 epochs
- Plug in validation data in each saved model to get predicted results
- Calculate NDCG@k and MAP@k

		NDCG@3	NDCG@5	NDCG@10	MAP@3	MAP@5	MAP@10	Time(s)	Model Size
Epoch=10	LSTM	0.7081	0.7291	0.7467	0.9489	0.9001	0.8482	39	
	Resnet	0.8287	0.8399	0.8399	0.9596	0.9352	0.9004	122	
	BERT	0.8595	0.8716	0.8823	0.9514	0.9252	0.8943	1574	
Epoch=20	LSTM	0.7117	0.7322	0.75	0.9496	0.902	0.8501	79	
	Resnet	0.8319	0.8427	0.8427	0.9627	0.9396	0.9067	251	
	BERT	0.8745	0.885	0.8946	0.9535	0.9311	0.9034	3164	
Epoch=30	LSTM	0.7126	0.7333	0.7508	0.9487	0.9006	0.8495	119	27.5M
	Resnet	0.8338	0.8432	0.8432	0.9628	0.9422	0.9096	380	42.5M
	BERT	0.8824	0.892	0.9009	0.9565	0.9361	0.9098	4753	436M

04 Evaluation - Metrics

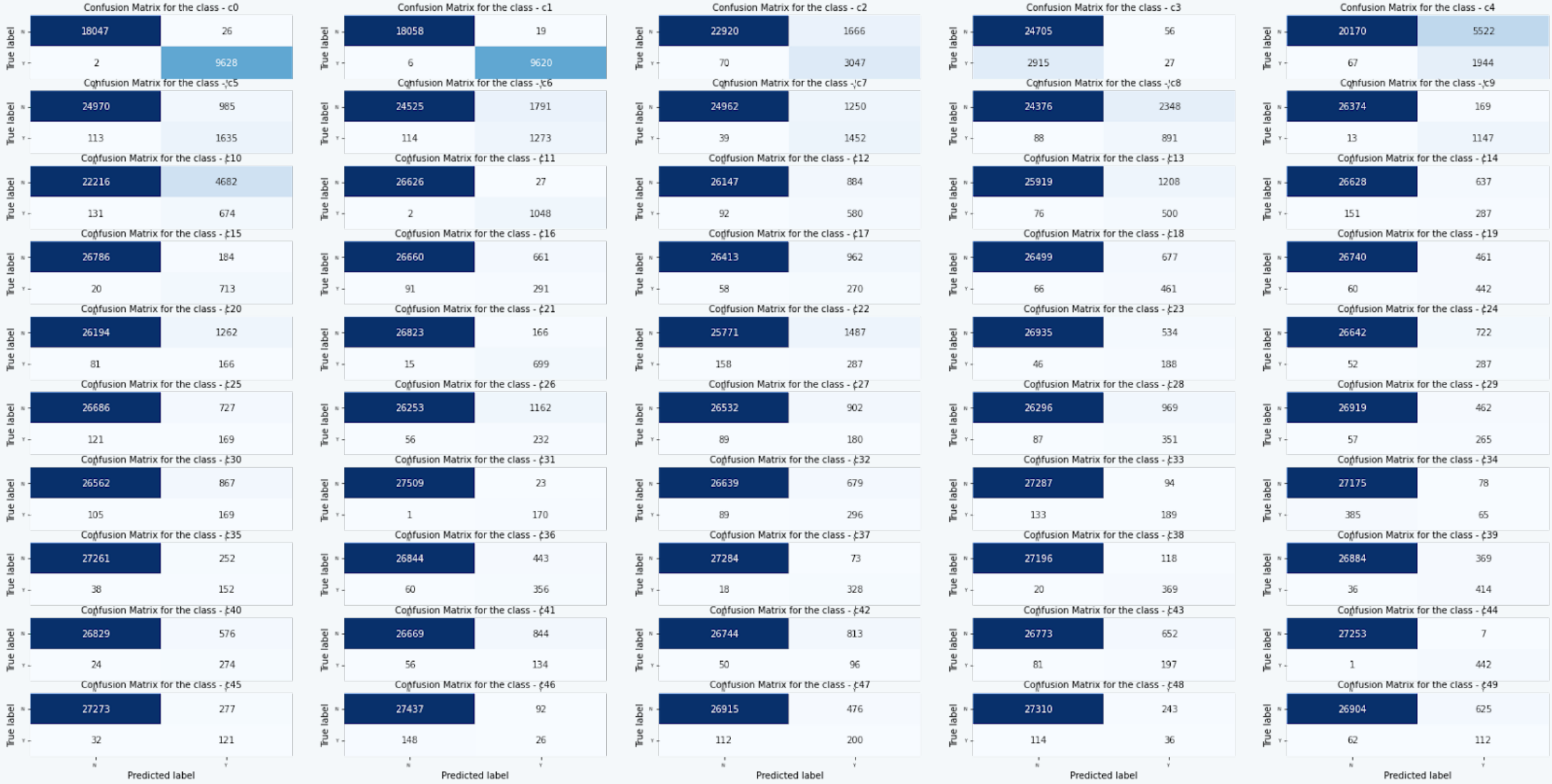
Presenter: Manxin Wang

- Compare f1 score, NDCG and MAP cross models
- TF-IDF has highest scores in NDCG and MAP; FastText has highest F1 score.

	F1 Score		NDCG			MAP		
	Micro	Macro	@3	@5	@10	@3	@5	@10
Random			0.3516	0.3799	0.4403	0.9850	0.8518	0.6556
TF-IDF + Logistic	0.65	0.51	0.9038	0.9109	0.9173	0.9681	0.9529	0.9338
FastText	0.66	0.55	0.8337	0.7379	0.6591	0.8587	0.6814	0.5659
LSTM	0.55	0.37	0.7126	0.7333	0.7508	0.9487	0.9006	0.8495
ResNet	0.62	0.48	0.8338	0.8432	0.8432	0.9628	0.9422	0.9096
Bert	0.64	0.52	0.8824	0.8920	0.9009	0.9565	0.9361	0.9098

04 Evaluation – TF-IDF + Logistic Confusion Matrix

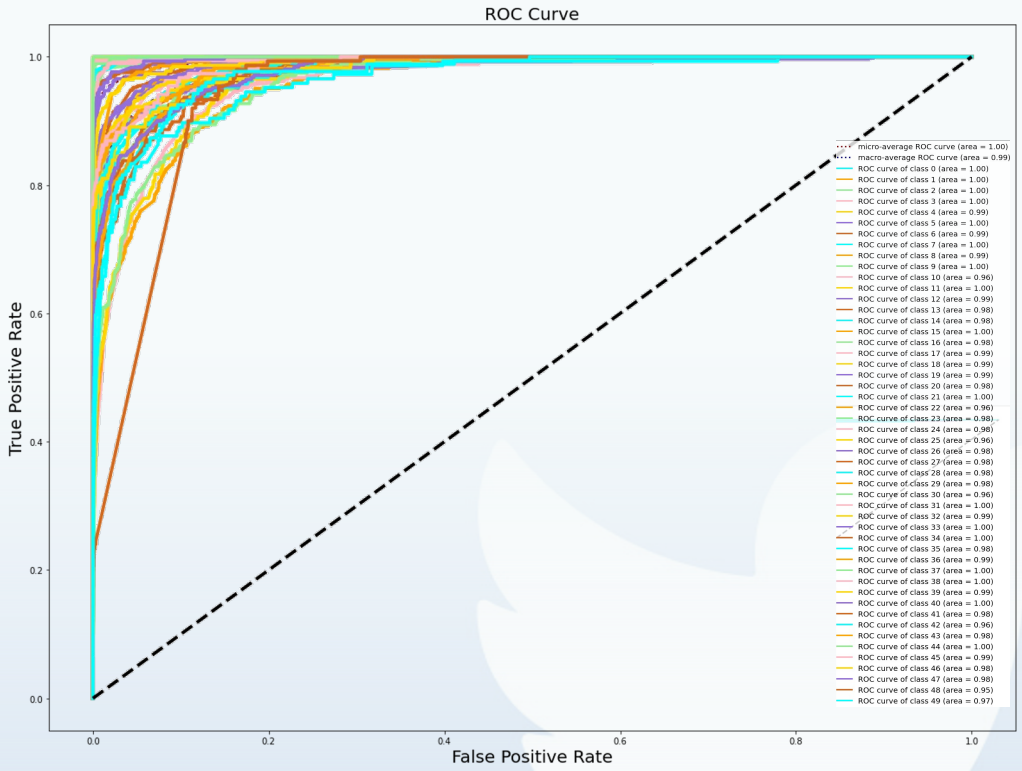
Presenter: Manxin Wang



04 Evaluation - ROC Curve & Precision and Recall Curve

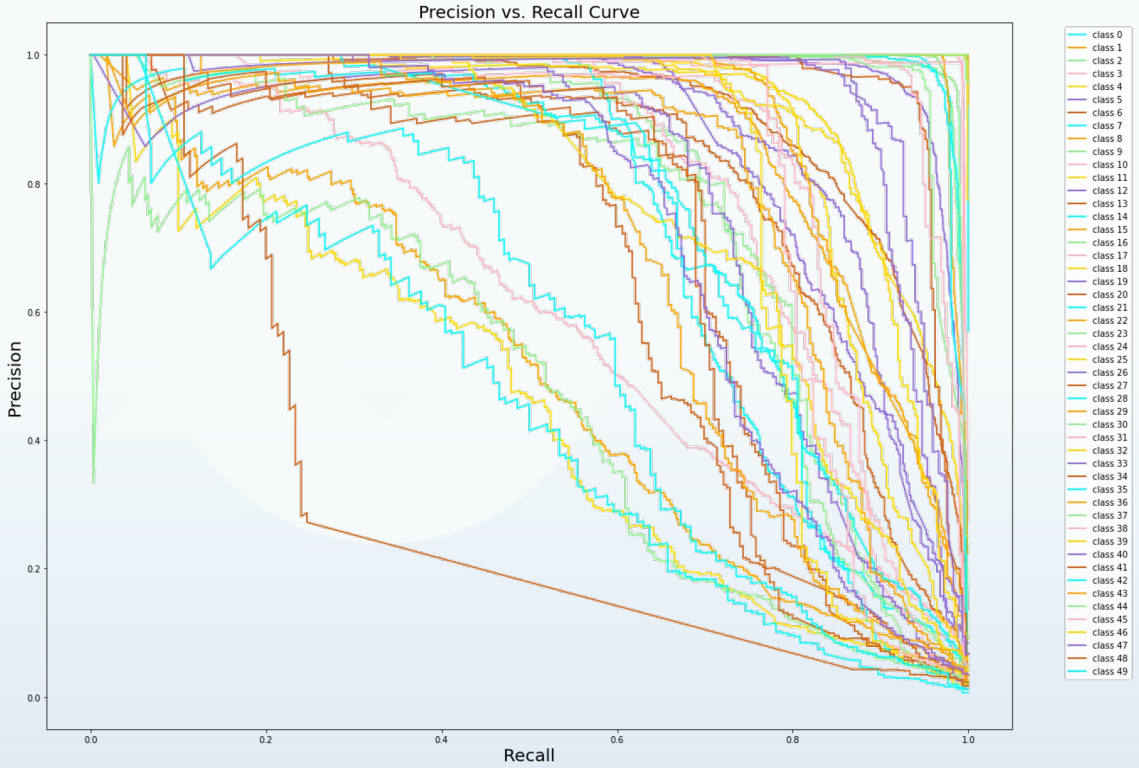
Presenter: Manxin Wang

TFIDF + Logistic Model ROC Curve



- The model overall ROC performance is outstanding
  AUC scores of top classes are close to one

TFIDF + Logistic Model Precision vs. Recall Curve



- Multi class imbalance shows in Precision and Recall curve
- The model do well in predicting the high frequency but not the less frequency

04 Evaluation – Statistical Evaluation

Presenter: Wanning Li

- Generate 50 samples from testset
- Predict each sample set and calculate NDCG and MAP scores
- Calculate 95% confidence interval to estimate population scores

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

	NDCG@3		NDCG@5		NDCG@10		MAP@3		MAP@5		MAP@10	
	Mean	Confidence Interval	Mean	Confidence Interval	Mean	Confidence Interval	Mean	Confidence Interval	Mean	Confidence Interval	Mean	Confidence Interval
Tfidf+Logistic	0.9038	(0.9004, 0.9073)	0.9109	(0.9077, 0.9141)	0.9173	(0.9144, 0.9202)	0.9681	(0.9662, 0.9699)	0.9529	(0.9510, 0.9549)	0.9338	(0.9312, 0.9364)
FastText	0.4801	(0.4745, 0.4857)	0.8247	(0.8201, 0.8292)	0.5605	(0.5557, 0.5654)	0.6665	(0.6615, 0.6716)	0.6373	(0.6336, 0.6411)	0.5426	(0.5380, 0.5472)
LSTM	0.7093	(0.7045, 0.7141)	0.7297	(0.7251, 0.7342)	0.7480	(0.7439, 0.7521)	0.9478	(0.9456, 0.9498)	0.9001	(0.8972, 0.9030)	0.8471	(0.8431, 0.8510)
ResNet	0.8356	(0.8321, 0.8391)	0.8450	(0.8416, 0.8483)	0.8561	(0.8529, 0.8594)	0.9653	(0.9635, 0.9671)	0.9447	(0.9424, 0.9469)	0.9120	(0.9091, 0.9149)
BERT	0.8841	(0.8806, 0.8877)	0.8939	(0.8907, 0.8971)	0.9027	(0.8997, 0.9056)	0.9591	(0.9571, 0.9612)	0.9383	(0.9358, 0.9408)	0.9128	(0.9101, 0.9155)

05 Error Analysis

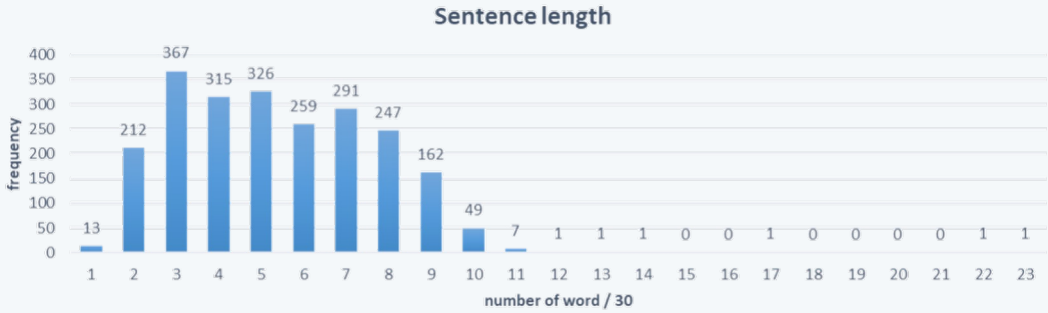
Presenter: Wanning Li

Predicted Results of ResNet

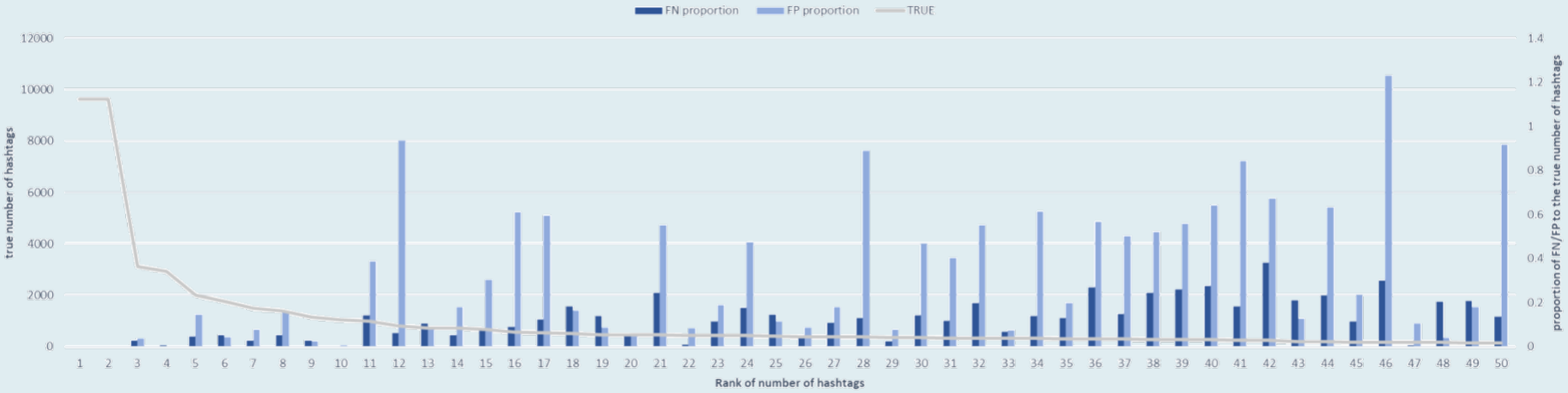
NDCG	Frequency
0.0 – 0.2	2254
0.2 - 0.4	1212
0.4 – 0.6	1023
0.6 – 0.8	888
0.8 – 1.0	22326

Data with NDCG between 0.0 – 0.2

Q: Is it because we only take the first 200 words of a sentence?
A: Not really.



Q: Which true hashtags were predicted Wrong (False Negative)? The hashtags(top 3 recommended) wrongly predicted (False Positive)?
A: The proportions of FN and FP are related to the size of dataset of hashtags.



Conclusion

Training data is not large enough to train a particularly good deep learning model!

06 Learnings and Future Work

Presenter: Wanning Li

Learnings

- Multilabel != Multiclass
- LSTM’s sentence embedding: pad at the beginning
- Dataset size should match model
- GPU limit for large deep learning model

Future Work

- Collect more data to provide better dataset
- Upgrade GPU limit to better tune parameter
- Use fastText word representation as word embedding
- Add function and design by streamlit cloud and deploy models

07 Demo – Twitter Hashtag Recommendation

Presenter: Huaiyue Chang

Paste tweet

Choose your model

- ResNet
- BERT
- LSTM
- fastText
- Logistics

Get N tags

3

1 10

Paste your 'Brand' related tweet below

Celebrate National STEM Day with a special virtual screening of “Not the Science Type” followed by a panel discussion moderated by @3M Chief Science Advocate Jayshree Seth featuring female STEM leaders and allies from our event sponsors. https://t.co/Wu2JgHjPuE

Get hashtags!

Hashtags Recommendation

Display Cleaned Tweet

celebrate national stem day special virtual screening science type follow panel discussion moderate chief science advocate jayshree seth feature female stem leader ally event sponsor

Tag Recommendation Results

	Hashtags	Probability
1	nowplaying	13.7%
2	ad	6.6%
3	nft	5.5%

Contribution

Data crawling	Wanning Li
Data cleaning for non DL models	Wanning Li, Yanan Zhou
Data cleaning for DL models	Qianran Ma, Yanan Zhou
Data preparation and description	Qianran Ma
Baseline - Random	Huaiyue Chang, Yanan Zhou
Tf-idf & Logistic Regression model	Yitong Wang, Yanan Zhou
FastText model	Huaiyue Chang
LSTM model	Wanning Li
Resnet model	Qianran Ma
BERT model	Yanan Zhou
Evaluation	Manxin Wang
Error analysis	Qianran Ma
Demo	Huaiyue Chang

Reference

[1] Zhang, Aston & Lipton, Zachary & Li, Mu & Smola, Alexander. (2021). Dive into Deep Learning.

[2] Winda Kurnia Sari, Rini DP, Reza Firsandaya Malik, Iman Saladin B. Azhar. Multilabel Text Classification in News Articles Using Long-Term Memory with Word2Vec. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) (Online). 2020;4(2):276-285. doi:10.29207/resti.v4i2.1655

[3] Joulin, Armand & Grave, Edouard & Bojanowski, Piotr & Mikolov, Tomas. (2016). Bag of Tricks for Efficient Text Classification.