Web Retrieval and Mining
Final Project

Sponsored Posts Recognizer for Restaurant Review

鄭筱樺, 王藝霖，林煦恩, 曹峻寧

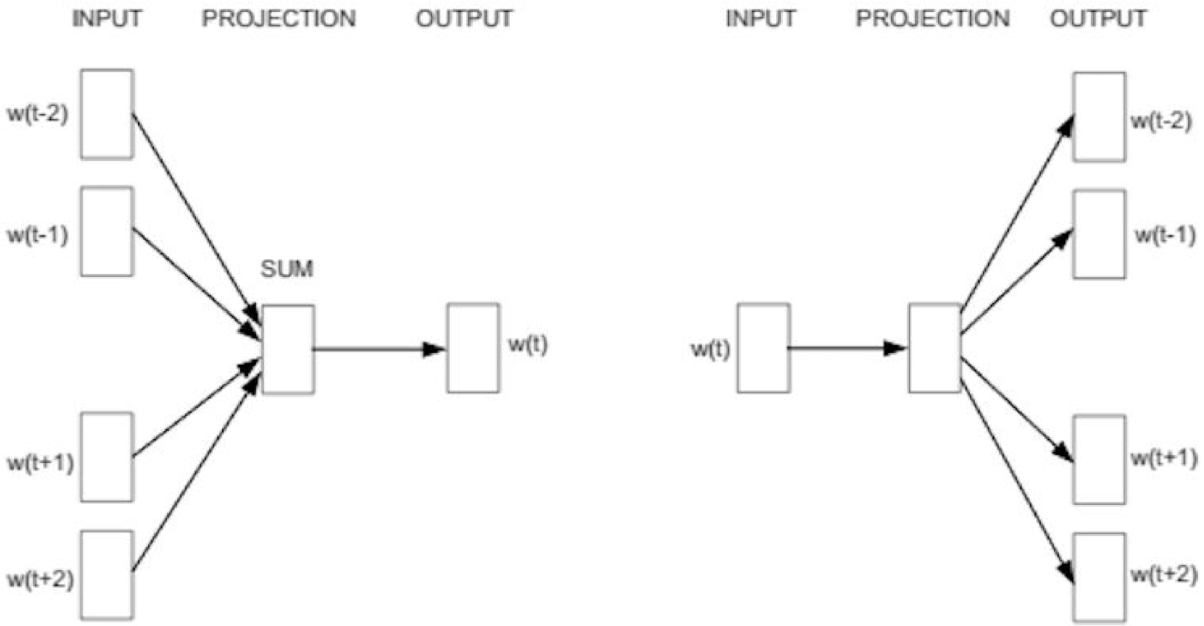
Outline

- Abstraction
- Motivation
- Methodology
  - word2vec
  - doc2vec
  - tfidf
  - linguistic-visual hybrid features
- Experiments
  - data set
  - evaluation
  - word2vec
  - ..
- Conclusion

Abstraction and Motivation

- sponsored posts recognizer for restaurant review
- what is sponsored post
    - a review post sponsored by the restaurant owner as an advertisement
- sponsored posts pretend to be a normal review
- be mislead and then visit those overvalued restaurants
- hard to recognized

Methodology - word2vec

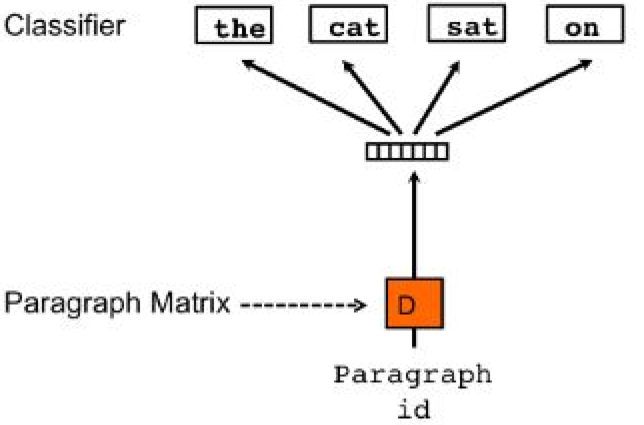


CBOW

Skip-gram

Methodology - doc2vec

- PV_DBOW



Methodology - TFIDF

Select features by document frequency

$$TF(t, d) = f_{t,d} \,/\, number\ of\ words\ in\ d$$

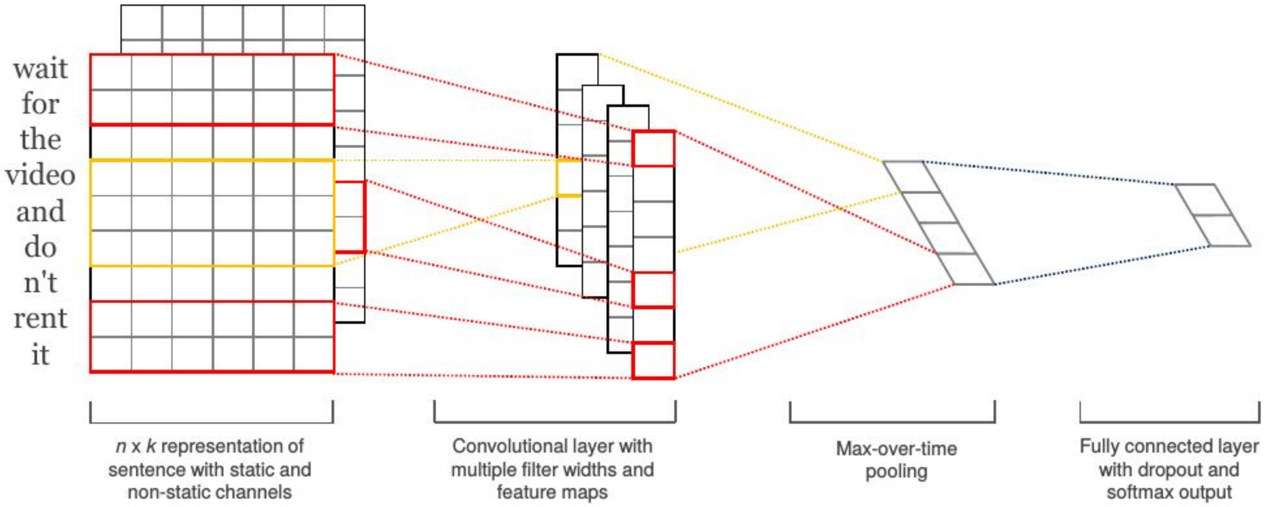
$$IDF(t) = log\ N \,/\, |\{d \in D :\ t \in d\}|$$

Methodology - Linguistic-Visual Hybrid Features (½)

- post content including text and image
- visual features
  - Human Faces
    - Haar feature-based cascade classifiers
  - Image Sharpness
    - the variation of the Laplacian by Pech-Pacheco et al

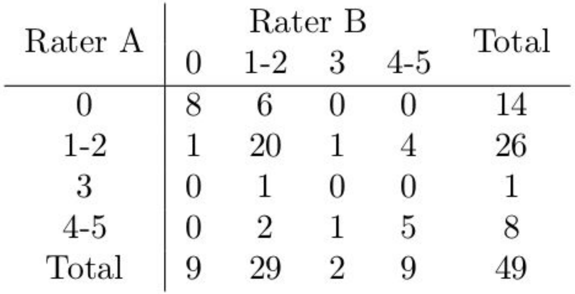
Methodology - Linguistic-Visual Hybrid Features (2/2)

- model


wait
for
the
video
and
do
n't
rent
it
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Experiments - Data Set

- PTT - training
  - by the rule of food board, the data is well labeled
- Blog (PIXNET) - testing
  - unlabeled
  - labeled by ourselves
  - kappa value =0.48 (which is an accepted value)



Rater A	Rater B				Total
	0	1-2	3	4-5	
0	8	6	0	0	14
1-2	1	20	1	4	26
3	0	1	0	0	1
4-5	0	2	1	5	8
Total	9	29	2	9	49

Experiments - Evaluation

1) *Cross Entropy Loss*

$$p \ = \ (a,\ 1 - a),\ q \ = \ (b,\ 1 - b)$$

$$H(p, q) =- \Sigma_i p_i log q_i$$

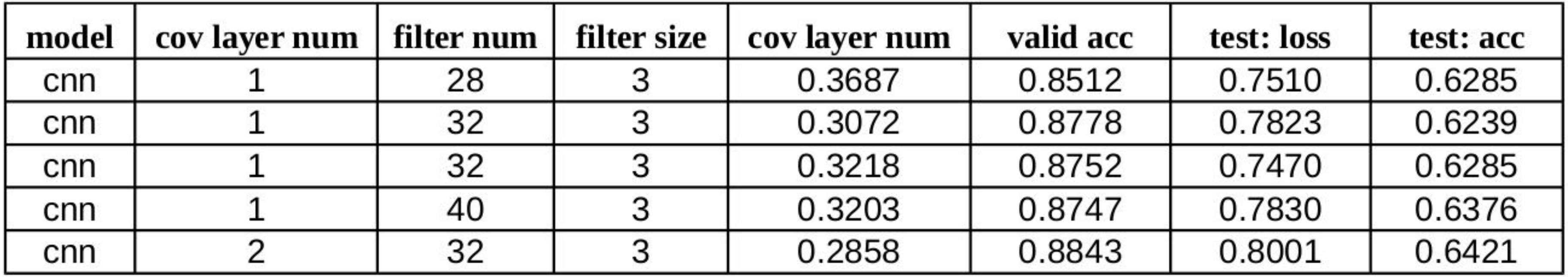
$$=- alogb - (1 - a)log(1 - b)$$

2) *Accuracy*

$$acc \ = \ \frac{number\ of\ correctly\ predicted\ posts}{number\ of\ ptt\ posts}$$

Experiments - Word2Vec

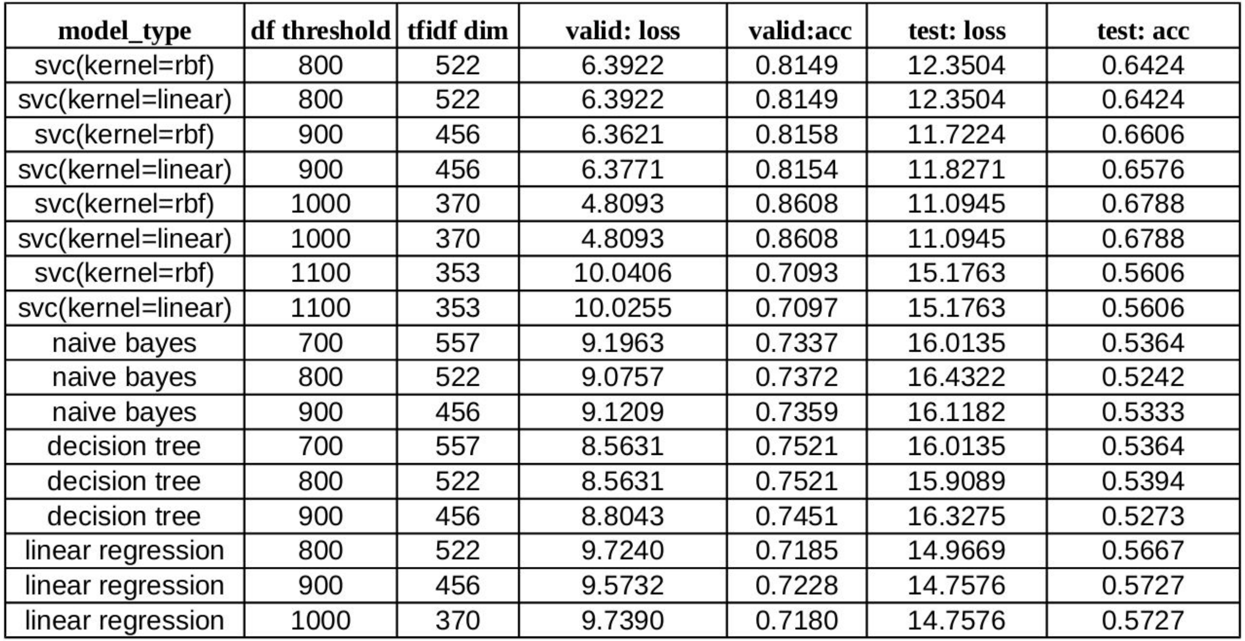
- 1 layer of convolutional is enough to generate good result



model	cov layer num	filter num	filter size	cov layer num	valid acc	test: loss	test: acc
cnn	1	28	3	0.3687	0.8512	0.7510	0.6285
cnn	1	32	3	0.3072	0.8778	0.7823	0.6239
cnn	1	32	3	0.3218	0.8752	0.7470	0.6285
cnn	1	40	3	0.3203	0.8747	0.7830	0.6376
cnn	2	32	3	0.2858	0.8843	0.8001	0.6421

Experiments - TFIDF

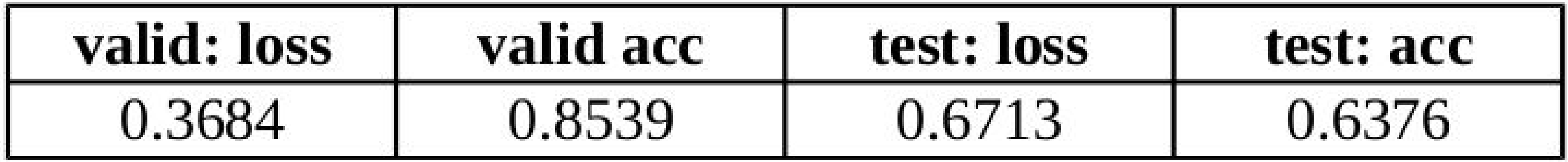
- svc with document frequency threshold 800 performs the best



model_type	df threshold	tfidf dim	valid: loss	valid:acc	test: loss	test: acc
svc(kernel=rbf)	800	522	6.3922	0.8149	12.3504	0.6424
svc(kernel=linear)	800	522	6.3922	0.8149	12.3504	0.6424
svc(kernel=rbf)	900	456	6.3621	0.8158	11.7224	0.6606
svc(kernel=linear)	900	456	6.3771	0.8154	11.8271	0.6576
svc(kernel=rbf)	1000	370	4.8093	0.8608	11.0945	0.6788
svc(kernel=linear)	1000	370	4.8093	0.8608	11.0945	0.6788
svc(kernel=rbf)	1100	353	10.0406	0.7093	15.1763	0.5606
svc(kernel=linear)	1100	353	10.0255	0.7097	15.1763	0.5606
naive bayes	700	557	9.1963	0.7337	16.0135	0.5364
naive bayes	800	522	9.0757	0.7372	16.4322	0.5242
naive bayes	900	456	9.1209	0.7359	16.1182	0.5333
decision tree	700	557	8.5631	0.7521	16.0135	0.5364
decision tree	800	522	8.5631	0.7521	15.9089	0.5394
decision tree	900	456	8.8043	0.7451	16.3275	0.5273
linear regression	800	522	9.7240	0.7185	14.9669	0.5667
linear regression	900	456	9.5732	0.7228	14.7576	0.5727
linear regression	1000	370	9.7390	0.7180	14.7576	0.5727

Experiments - Linguistic-Visual Hybrid Features

The result is not so well



valid: loss	valid acc	test: loss	test: acc
0.3684	0.8539	0.6713	0.6376

Conclusion

- Result Discussion
  - word2vec outperform other methodologies
- Future Work
  - Data Labeling
    - ambiguous label
    - more labeled data
  - Extended Application
    - recognize “wrong” information from a set
    - fake news recognition or content farm detection