Predicting Instructor’s Intervention in MOOC forums

Snigdha Chaturvedi, Dan Goldwasser, Hal Daume IIII

ACL 2014

2019.03.29

1
BAC

研究背景

目的：

预测在MOOC的论坛中，某个帖子是否需要指导老师干预。--二分类问题

必要性：

- MOOC现在非常流行，其课程质量高且比较费用较低
- 缺少个性化的教育体验
    - 和指导老师的互动主要依靠发帖实现
    - 课程学生多，老师少，课程时间短，指导老师很难参与所有帖子。

1 BAC

研究背景

假设：

- 一般而言，导师需要在临近测验或者考试的时候需要介入，他们提问的问题可能具有一些暗示作用，比如：关于分数等
- 如果学生非常想要导师回答某个问题，可能通过一种“up-voting”的方式

“The problem summary: Anyone else having problems viewing the video lecture...very choppy. If you are also experiencing this issue; please upvote this post.”
“I read that by up-voting threads and posts you can get the instructors’ attention faster.”
“Its is very bad to me that I achieved 10 marks in my 1st assignment and now 9 marks in my 2nd assignment, now I won’t get certificate, please Course staff it is my appeal to change the passing scheme or please be lenient. Please upvote my post so that staff take this problem under consideration.”

Figure 1: Sample posts that showing students desiring instructor’s attention have to resolve to the inefficient method of getting their posts upvoted.

1
BAC

研究背景

解决方案：

- Logistic regression model using high level information about threads and posts.
- Linear Chain Markov Model
- Global Chain Model

- Individual posts belong to latent categories which represent their textual content at an abstract level
- Instructor’s decision to reply to a post is based on chain of events.

2 Pro

问题设置

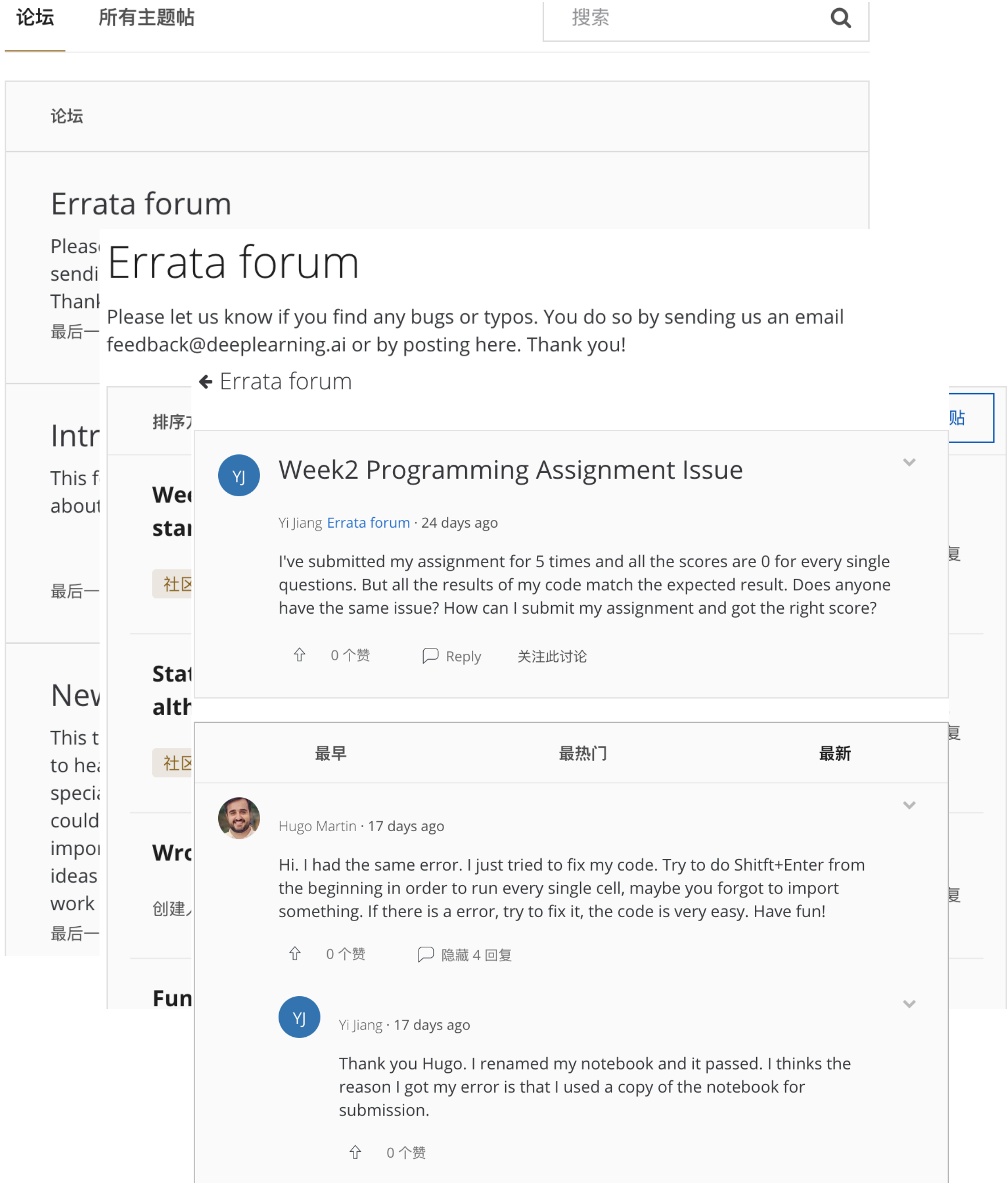
Forum：代表“主题”，“Assignment”、“Study Group”

Thread： A Forum consists of multiple threads, each thread (t) has a title.

Post: A Thread consists of multiple posts (p_i).

Decision：表示为r，若需要指导老师回复某thread，则值为1，否则为0.

单独的Post没有title，每个thread的post数目都是不一样的。



2 Mod

模型-LR

模型目标： $p(r|t)$

使用了2个类型的Feature： *Thread only features*、*Aggregated post features*

capture information about the thread such as when, where, by who and lexical features

compress the features from posts

2 Mod

模型-LR

Thread only features:

1. a binary feature indicating if the thread was started by an anonymous user
2. three binary features indicating whether the thread was marked as approved, unresolved or deleted (respectively)
3. forum id in which the thread was posted
4. time when the thread was started
5. time of last posting on the thread
6. total number of posts in the thread
7. a binary feature indicating if the thread title contains the words *lecture* or *lectures*
8. a binary feature indicating if the thread title contains the words *assignment*, *quiz*, *grade*, *project*, *exam* (and their plural forms)

Aggregated post features:

9. sum of number of votes received by the individual posts
10. mean and variance of the posting times of individual posts in the thread
11. mean of time difference between the posting times of individual posts and the closest course landmark. A course landmark is the deadline of an assignment, exam or project.
12. sum of count of occurrences of assessment related words e.g. *grade*, *exam*, *assignment*, *quiz*, *reading*, *project* etc. in the posts
13. sum of count of occurrences of words indicating technical problems e.g. *problem*, *error*
14. sum of count of occurrences of thread conclusive words like *thank you* and *thank*
15. sum of count of occurrences of *request*, *submit*, *suggest*

2
Mod

模型-LR

也考虑了一些其他的特征，比如：who started the thread, number of distinct participants in the thread,…

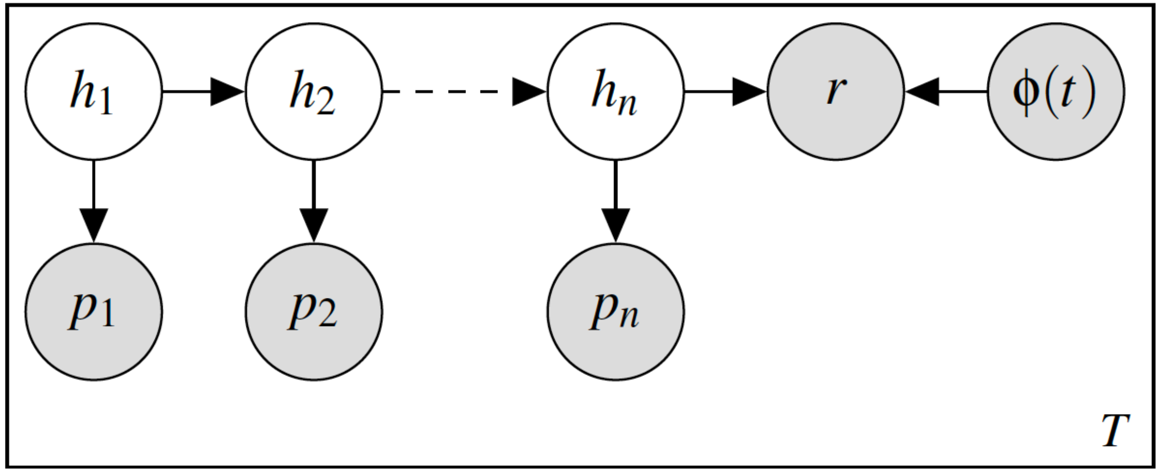
但是模型结果表明，这些特征是无效的。

2 Mod

模型-LCMM

LR缺点：

- LR模型可以很好地挖掘thread层面的特征，但是不能很好地挖掘单个post的特征
- 通过*Aggregated post features*可以挖掘到一些post信息，但是因为每个thread中包含的post数目不定，所以post的特征依赖于Aggregated的值
- 之前设计的特征，post之间是没有相互依赖的，但实际上，他们是有时间顺序的



(a) Linear Chain Markov Model (LCMM)

Score function of a thread：

$$f_w(t,p) = \max_h [\mathbf{w} \cdot \boldsymbol{\phi}(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{h}, \boldsymbol{t})]$$

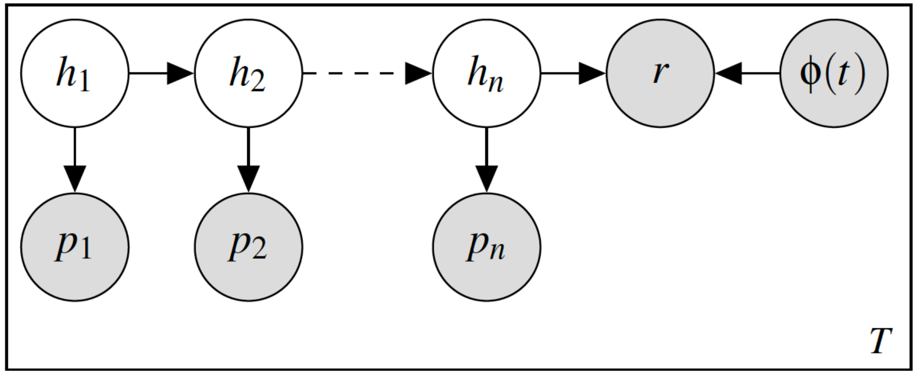
2 Mod

模型-LCMM

For every thread, t, in the dataset:

1. Choose a start state, h_1, and emit the first post, p_1.
2. For every subsequent post, $p_i\ \forall\ i \in \{2 \ldots n\}$:
   (a) $\boxed{\text{Transition from } h_{i-1} \text{ to } h_i.}$
   (b) Emit post p_i.
3. Generate the instructor's intervention decision, r, using the last state h_n and non-structural features, $\phi(t)$.

Figure 3: Instructor's intervention decision process for the Linear Chain Markov Model.



(a) Linear Chain Markov Model (LCMM)

Algorithm 1 Training algorithm for LCMM

Input: Labeled data $D = \{(t, p, r)_i\}$
Output: Weights w
Initialization: Set w_j randomly, $\forall j$
for $t : 1$ to N **do**
viterbi
    $\boxed{\hat{h}_i = \arg\max_h [\mathbf{w_t} \cdot \boldsymbol{\phi}(\boldsymbol{p}, \boldsymbol{r}, \boldsymbol{h}, \boldsymbol{t})]}$ such that $r = r_i \forall i$
    w_{t+1} = StructuredPerceptron$(t, p, \hat{h}, r)$
end for
return w

Collins M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 1-8.

模型-LCMM

Post Emission Features:

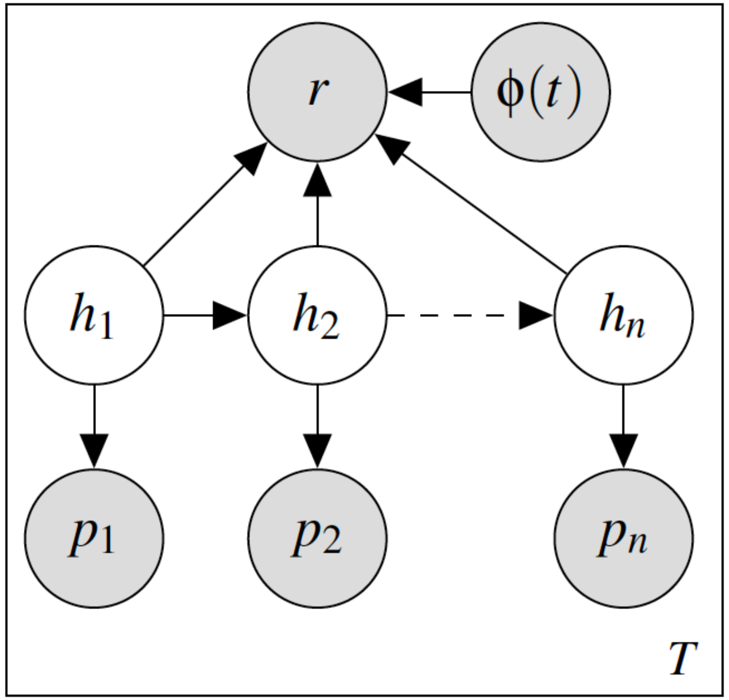
1. $\phi(p_i, h_i)$ = count of occurrences of question words or question marks in p_i if the state is h_i; 0 otherwise.
2. $\phi(p_i, h_i)$ = count of occurrences of thank words (*thank you* or *thanks*) in p_i if the state is h_i; 0 otherwise.
3. $\phi(p_i, h_i)$ = count of occurrences of greeting words (e.g. *hi*, *hello*, *good morning*, *welcome* etc) in p_i if the state is h_i; 0 otherwise.
4. $\phi(p_i, h_i)$ = count of occurrences of assessment related words (e.g. *grade*, *exam*, *assignment*, *quiz*, *reading*, *project* etc.) in p_i if the state is h_i; 0 otherwise.
5. $\phi(p_i, h_i)$ = count of occurrences of *request*, *submit* or *suggest* in p_i if the state is h_i; 0 otherwise.
6. $\phi(p_i, h_i)$ = log(course duration/$t(p_i)$) if the state is h_i; 0 otherwise. Here $t(p_i)$ is the difference between the posting time of p_i and the closest course landmark (assignment or project deadline or exam).
7. $\phi(p_i, p_{i-1}, h_i)$ = difference between posting times of p_i and p_{i-1} normalized by course duration if the state is h_i; 0 otherwise.

模型-LCMM

Transition Features:

1. $\phi(h_{i-1}, h_i) = 1$ if previous state is h_{i-1} and current state is h_i; 0 otherwise.
2. $\phi(h_{i-1}, h_i, p_i, p_{i-1}) =$ cosine similarity between p_{i-1} and p_i if previous state is h_{i-1} and current state is h_i; 0 otherwise.
3. $\phi(h_{i-1}, h_i, p_i, p_{i-1}) =$ length of p_i if previous state is h_{i-1}, p_{i-1} has non-zero question words and current state is h_i; 0 otherwise.
4. $\phi(h_n, r) = 1$ if last post's state is h_n and intervention decision is r; 0 otherwise.
5. $\phi(h_n, r, p_n) = 1$ if last post's state is h_n, p_n has non-zero question words and intervention decision is r; 0 otherwise.
6. $\phi(h_n, r, p_n) = \log(\text{course duration}/t(p_n))$ if last post's state is h_n and intervention decision is r; 0 otherwise. Here $t(p_n)$ is the difference between the posting time of p_n and the closest course landmark (assignment or project deadline or exam).

模型-GCM



(b) Global Chain Model (GCM)

$$\min_{w} \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_{j}^{T} l(-r_j f_w(t_j, p_j))$$

$square\ hinge\ loss: \max(0, 1 - r_j f_w(t_j, p_j))^2$

模型-GCM

$$\min_{w} \frac{\lambda}{2}||\mathbf{w}||^2 + \sum_{j}^{T} l(-r_j f_w(t_j, p_j))$$

Algorithm 1 *LCLR* :The algorithm that optimizes (3)

1: initialize: $\mathbf{u} \leftarrow \mathbf{u}_0$
2: **repeat**
3: **for all** positive examples $(\mathbf{x}_i, y_i = 1)$ **do**
4: Find $\mathbf{h}_i^* \leftarrow \arg\max_{\mathbf{h}\in\mathcal{C}} \sum_s h_s \mathbf{u}^T \phi_s(\mathbf{x}_i)$
5: **end for**
6: Update $\mathbf{u}$ by solving

$$\min_{\mathbf{u}} \frac{\lambda}{2}\|\mathbf{u}\|^2 + \sum_{i:y_i=1} \ell(-\mathbf{u}^T \sum_s h_{i,s}^* \phi_s(\mathbf{x}_i)) + \sum_{i:y_i=-1} \ell(\max_{\mathbf{h}\in\mathcal{C}} \mathbf{u}^T \sum_s h_s \phi_s(\mathbf{x}_i)) \quad (4)$$

7: **until** convergence
8: return $\mathbf{u}$

Algorithm 2 Cutting plane algorithm to optimize Eq. (4)

1: for each negative example x_j, $H_j \leftarrow \emptyset$
2: **repeat**
3: **for** each negative example x_j **do**
4: Find $\mathbf{h}_j^* \leftarrow \arg\max_{\mathbf{h}\in\mathcal{C}} \sum_s h_s \mathbf{u}^T \phi_s(\mathbf{x}_j)$
5: $H_j \leftarrow H_j \cup \{\mathbf{h}_j^*\}$
6: **end for**
7: Solve

$$\min_{\mathbf{u}} \frac{\lambda}{2}\|\mathbf{u}\|^2 + \sum_{i:y_i=1} \ell(-\mathbf{u}^T \sum_s h_{i,s}^* \phi_s(\mathbf{x}_i)) + \sum_{i:y_i=-1} \ell(\max_{\mathbf{h}\in H_j} \mathbf{u}^T \sum_s h_s \phi_s(\mathbf{x}_i)) \quad (5)$$

8: **until** no new element is added to any H_j
9: return $\mathbf{u}$

Chang M W, Goldwasser D, Roth D, et al. Discriminative learning over constrained latent representations[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 429-437.

Experiments Results

DataSets:

①Genes and the Human Condition (From Behavior to Biotechnology) (GHC) dataset.

- 30,000 students.
- 2 instructors.
- 3 teaching assistants.
- 56 technical support staff.
- 980 threads composed of about 3,800 posts.

②Women and the Civil Rights Movement (WCR) dataset.

- 14, 600 students
- 1 instructor
- 6 teaching assistants
- 49 support staff
- 800 threads composed of 3,900 posts.

Test data

- GHC： 186 threads out of which the instructor intervened on 24 while.
- WCR： 155 threads out of which the instructor intervened on 21 while.

Experiments Results

Model	Genes and the Human Condition (GHC)			Women and the Civil Rights (WCR)		
	P	R	F	P	R	F
LR	44.44	16.67	24.24	66.67	15.38	25.00
J48	45.50	20.80	28.55	25.00	23.10	24.01
LCMM	33.33	29.17	31.11	42.86	23.08	**30.00**
GCM	60.00	25.00	**35.29**	50.00	18.52	27.03

Table 1: Held-out test set performances of chain models, LCMM and GCM, are better than that of the unstructured models, LR and J48.

Experiments Results

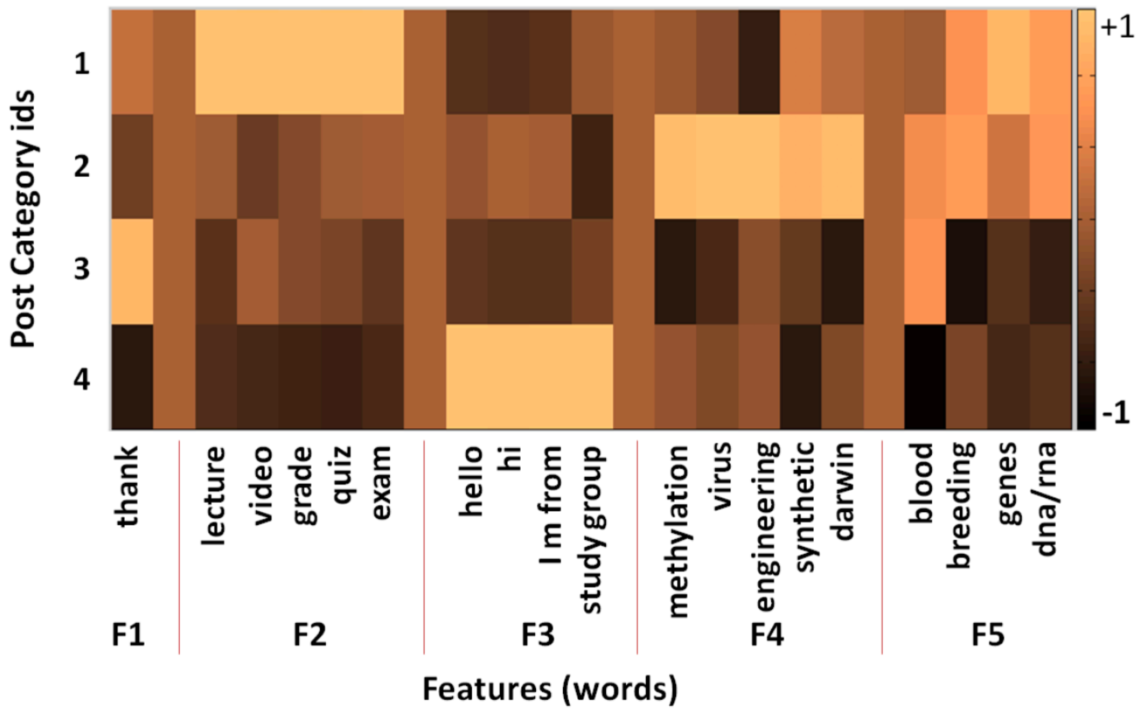


Figure 4: Visualization of lexical contents of the categories learnt by our model from the GHC dataset. Each row is a category and each column represents a feature vector. Bright cream color represents high values while lower values are represented by darker shades. Dark beige columns are used to better separate the five feature clusters, F1-F5, which represent words that are common in thanking, logistics-related, introductory, syllabus related and miscellaneous posts respectively. Categories 1,2,3 and 4 are dominated by F2, F4, F1 and F3 respectively indicating a semantic segregation of posts by our model's categories.

3 EXP

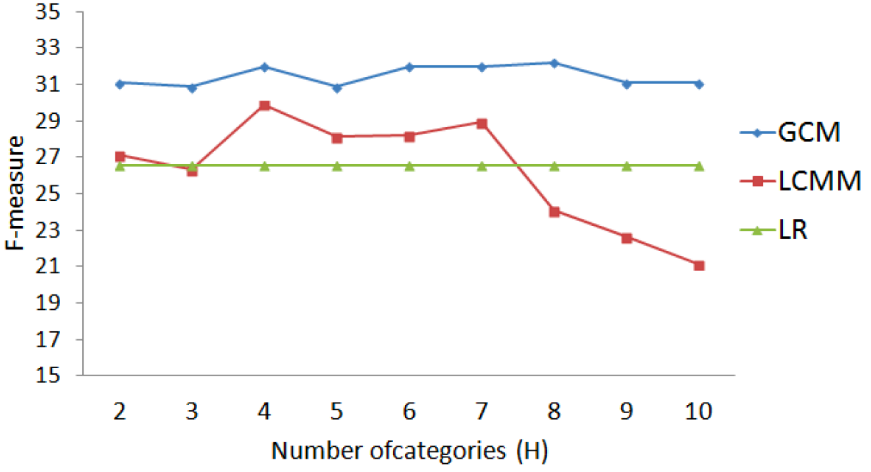
Experiments Results

Category	Example posts
1	'I'm having some issues with video playback. I have downloaded the videos to my laptop...'
1	'There was no mention of the nuclear envelope in the Week One lecture, yet it was in the quiz. Is this a mistake?'
2	'DNA methylation is a crucial part of normal development of organisms and cell differentiation in higher organisms...'
2	'In the lecture, she said there are...I don't see how tumor-suppressor genes are a cancer group mutation.'
3	'Thank you very much for a most enjoyable and informative course.'
3	'Great glossary! Thank you!'
4	'Hello everyone, I'm ... from the Netherlands. I'm a life science student.
4	'Hi, my name is ... this is my third class with coursera'

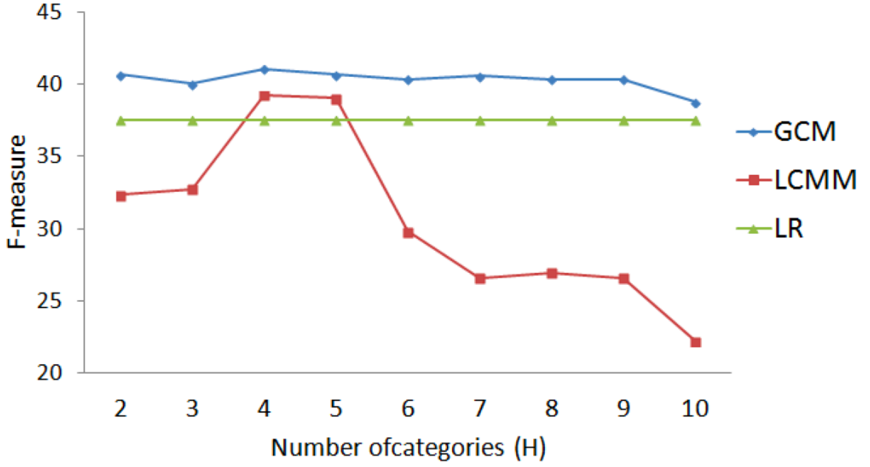
Table 2: Representative posts from the four categories learnt by our model. Due to space and privacy concerns we omit some parts of the text, indicated by "...".

3 EXP

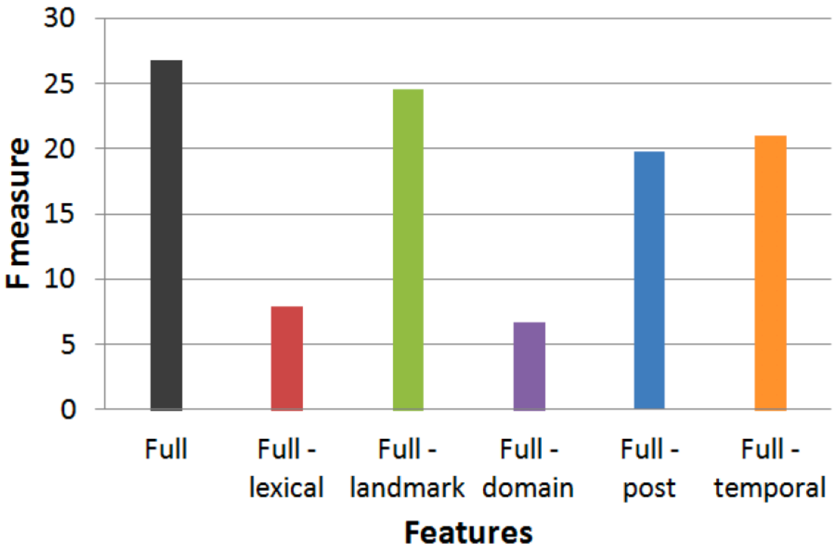
Experiments Results



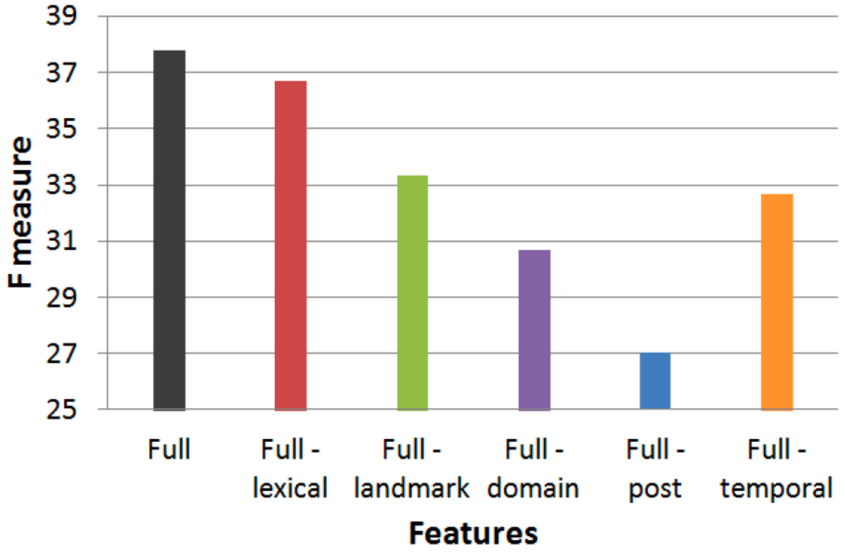
(a) Genes and the Human Condition dataset



(b) Women and the Civil Rights Movement dataset



(a) Genes and the Human Condition dataset



(b) Women and the Civil Rights Movement dataset

Figure 6: Cross validation performances of the various feature types for the two datasets.

lexical: title, posts
landmark: exams, quizzes
domain: lexical + landmark
post: only aggregated posts infor.
temporal: time started/ended

Conclusion

贡献点

- 结合帖子内容和帖子的行为特征
- 提出了3个模型解决这个问题

可改进点：

- 二分类问题感觉不是很合适，改为排序问题可能更合理
- 基于帖子内容、主题等挖掘太浅，同一帖子的内容顺序没有考虑，转折等连接词没有（情感可以融入等）
- 涉及知识点的难易程度

Thanks!