Modeling Patient Mortality from Clinical Text by Combining Topic Modeling and Ontological Feature Learning with Group Regularization

Jeong Min Lee[1], Charmgil Hong[2], and Milos Hauskrecht[3]

Abstract— **In this project, we modeled mortality of Intensive Care Unit (ICU) patient by unstructured clinical notes. The mortality of ICU patient is critical for the betterment of patient care since it provides the summarization of a patient's severity from complex physiological information. Our objective is to make a model that learns a compact representation of clinical text feature space consisted of two different text-derived features; semantically enriched concepts from ontology and topics generated from probabilistic topic modeling. The underlying assumption is that a feature space of richer information can be obtained when the two different feature learning schemes are combined. With these various feature sets, we find a compact representation and classification by sparse group regularization on logistic regression. This representation is more compact and leads better predictability on patient mortality.**

I. INTRODUCTION

Mortality of ICU not only indicates likeliness of death of a patient. It is an abstracted signal that tells how much a patient is in a severe situation from complex physiological signals and information can be observed and gathered in ICU. Several mortality scoring systems have been devised and widely used, the sources of those measurements are only based on the structured physiological features such as heart rate and temperature. Clinical progress note which is unstructured free hand-note written by physicians and nurses contains information about the physiology of patient in detail. Unlike its importance, most ICU mortality modeling uses primarily structured data and physiological waveforms [1].

Developing predictive models with unstructured text is a challenging task since it resides in high dimensional bag-of-word vector space. Our goal is to make a model that learns a compact representation of clinical text feature space consisted of two different text-derived features; semantically enriched concepts from ontology and topics generated from probabilistic topic modeling. The underlying assumption is that a feature space of richer information can be obtained when the two different feature learning schemes are combined. The underlying assumption of this approach is that a feature space of richer information can be obtained when the two different feature learning schemes, probabilistic dimensionality reduction, and ontology-based feature transformation, are combined. A probabilistic dimensionality reduction approach, Latent Dirichlet Allocation (LDA) [2], produces bag-of-topic representation and knowledge-based feature transformation process derives bag-of-concept feature set from the term mapping with a biomedical ontology. These two additional feature groups provide supplementary information which original bag-of-word feature might not reveal. Especially, the concepts reaped from ontology encodes expert's various clinical knowledge. With these various feature sets, we find a compact representation and classification by sparse group regularization on logistic regression.

[1]Jeong Min Lee is with Department of Computer Science, University of Pittsburgh, 210 S Bouquet St. Pittsburgh, PA 15260, USA jlee@cs.pitt.edu

[2]Charmgil Hong is with Department of Computer Science, University of Pittsburgh, 210 S Bouquet St. Pittsburgh, PA 15260, USA charmgil@cs.pitt.edu

[3]Milos Hauskrecht is with Department of Computer Science, University of Pittsburgh, 210 S Bouquet St. Pittsburgh, PA 15260, USA milos@cs.pitt.edu

II. RELATED WORKS

A. Topic Modeling in Clinical Domain

Topic modeling which also referred to Latent Dirichlet Allocation is a generative method that creates hidden topics from observable document consisted of word counts multinominal data. It assumes the topics as Dirichlet random variable and assigns each word to topics as the multinomial random variable. Then, a document which exhibits multiple topics is modeled as Dirichlet random variable of topics. Since its capability of capturing meaningful structure in the data, it has been widely used in many applications. With regard clinical domains, it has been used as a method of dimensionality reduction for classification tasks and also finding a hidden structure of data. For dimensionality reduction purpose, it has been used to predict patient satisfaction [3], depression [4], infection [5], and mortality [6].

B. Feature Selection using Ontology

Ontology is a formal representation of the concepts and its relationship of specific domain or area. In practical perspective, ontology can be regarded as a database of a specific area that provides a description of concepts and relationships of concepts. In biomedical area, there exist ontologies that covering specific and general biological and medical concepts such as medications, lab tests, diseases, and genes. With a plethora of expert-curated biomedical ontologies, it has been used to phenotype disease using human gene ontology [7], heart failure readmission prediction on drug data using drug ontology [8], and gene expression classification [9]. Many utilization of ontologies on feature selection has been dominated by genetic, proteomic, and drug-related area and not many attempts have been made within patient mortality prediction on clinical note data.

III. APPROACH

In overall, the process of mortality modeling of a patient from unstructured clinical notes involves two large steps. First, we generated derivative feature sets after creating the bag-of-word feature set from the unstructured free text. Second, we modeled the mortality by casting it classification problem with regularizations.

A. Feature Generation

First of all, we created word feature from the unstructured free text and it is the baseline of our feature combinations.

Regarding dataset, we used MIMIC-3 Database [10], publicly available de-identified medical record in critical care dataset. It contains medical records of around 46,000 patients who hospitalized in Beth-Israel Deaconess hospital between 2001 and 2012. We created labels regarding whether a patient died while in the first hospital admission (positive class) or not (negative class) and that resulted in a high class-imbalance problem: only 10.3 percent patients died in the first hospital admission. To resolve the issue, we subsampled on removing negative class sample randomly such that the ratio of the positive and negative class is 1:2. We also limited patient to be between the age of 18 and 99 and the notes to be categories of nursing and physician notes. We removed all notes written on the day of death or discharge. The resulting samples then consist of 5334 patients.

1) Word Feature: In order to create the word feature, we merged all notes of a patient in the first admission and represent each note as bag-of-word feature. As a preprocessing step, word stemming and removing stopword processed with a standard stopword list provided in R package tm. Also, we removed words that occurred less than 48 times. The resulting words are 10229 and represented as the bag-of-word representation.

2) Concept Feature: Based on the word feature, we create two derivative features: the concept feature and topic feature. The concept feature consists of clinical text matched concepts from a clinical ontology. For the matching process, UMLS (Unified Medical Language System) concept-text mapping framework, Meta Map, is used to find concept from SNOMED-CT ontology. The Meta Map finds matched concept with following steps and details can be found in [11]:

- (1) Parse note text into noun phrases.
- (2) Generate variants of each phrases using the SPECIALIST lexicon which specialized in biomedical vocabularies and maintained by NLM. The variants include acronyms, abbreviations, synonyms, meaningful combinations, etc.
- (3) Retrieve candidate sets includes all ontological concepts containing at least one of the variants.
- (4) Evaluate candidates against the input text by first computing a mapping from the phrase words to the candidate concepts and then evaluating the strength of the mapping using an evaluation function of weighted sum of four metrics: *centrality* (involvement of the head), *variation* (an average of inverse distance scores), *coverage* and *cohesiveness*, which measures how much of a candidate matches the text and in how many pieces. Then candidate concepts sorted according to mapping score.

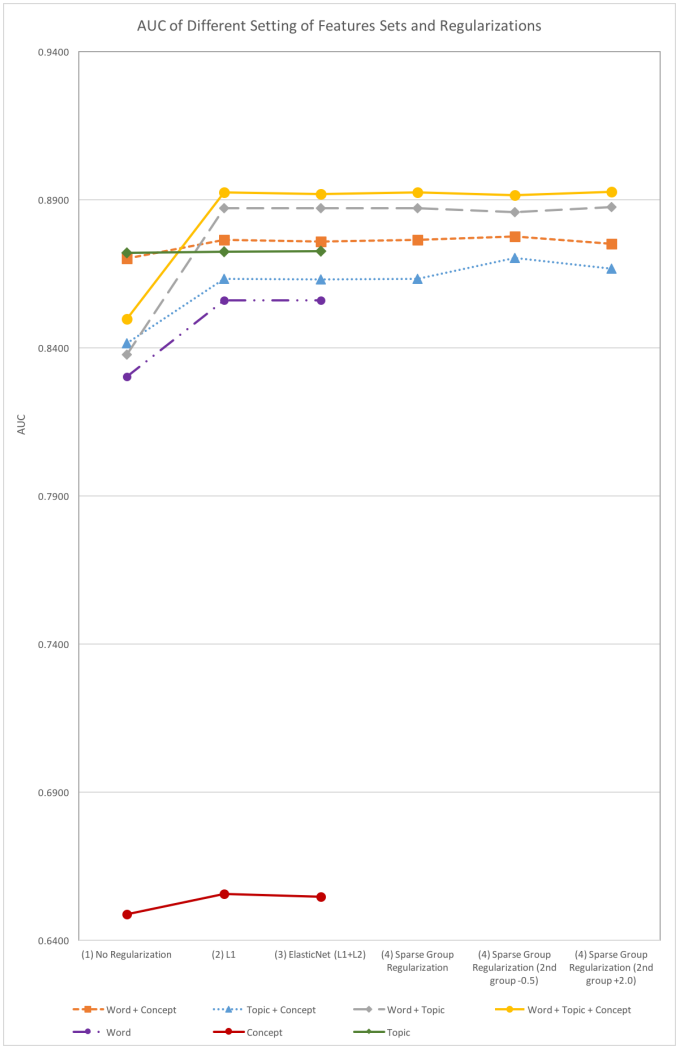

Fig. 1. AUC (Area Under Receiver Operating Curve) of Different Setting of Feature Sets and Regularizations

We restrict matching concept to be in specific semantic types: *Age group, anatomical structure, antibiotic, body part, organ, or organ component, clinical drug, diagnostic procedure, disease or syndrome, finding, organism, pharmacologic substance, physiologic function, and, sign or symptom*. We remove concepts occurred less than 20 times and resulting number of concept feature is 118.

3) Topic Feature: We created the topic feature using LDA. The process of LDA can be seen as projecting the word feature to lower dimensional latent space and we call the features in that space as bag-of-topics. LDA is a generative, bayesian model that best known as finding hidden topics from a document of a corpus. The generative probabilistic model assumes that observed random variable is generated from some probability distribution so that its main task is to know the type and estimate parameter of the distribution.

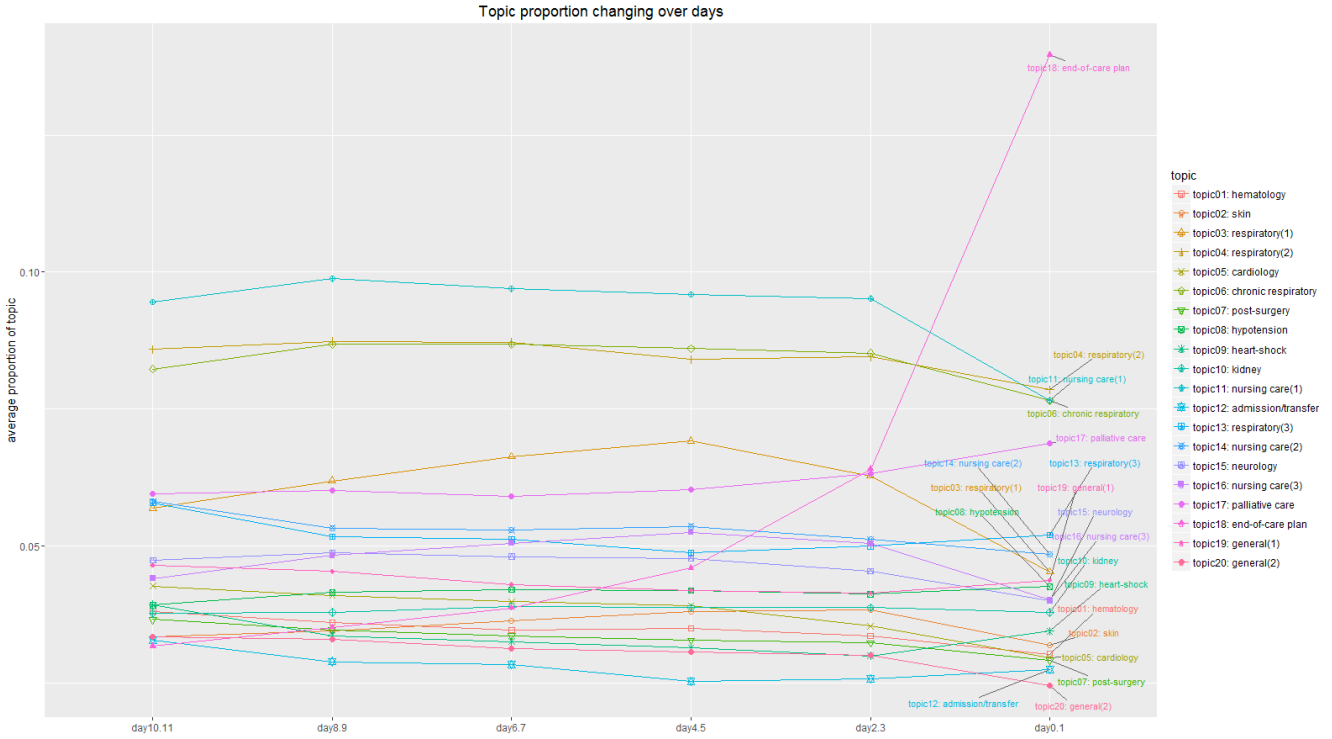

Fig. 2. Topic Proportion Changing over Last 10 Days

As the generative model, LDA models topics as Dirichlet random variable and assignment of each word to a topic as the multinomial random variable. The document which exhibits multiple topics is modeled as Dirichlet random variable of topics. For the topic modeling in our work, the number of topics is set as 50 since our previous preliminary study showed no improvement in classification with more than 50 topics. The topic modeling process involving Gibbs Sampling with 300 chains.

4) Mixed Feature Sets: We create mixed feature sets by combining the three feature sets we generated so far.

- Topic + Concept feature (Dim: 168)
- Word + Concept feature (Dim: 10318)
- Word + Topic feature (Dim: 10279)
- Word + Topic + Concept feature (10379)

B. Feature Regularization and Classification

With these various feature sets in hand, we find a compact representation by sparse group regularization on logistic regression. We tried different regularization methods to see their effect in classification. Note that the objective functions are the negative log-likelihood loss with a class encoding of -1 and +1.

1) L1 Regularization: L1 regularization selects features by shrinking coefficient of certain features while solving linear square problem.

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i(\mathbf{x}_i^T\mathbf{w}+c)))+\lambda\|\mathbf{w}\|_1$$

where $\|\mathbf{w}\|_1 = \sum_j |w_j|$

2) Elastic Net Regularization: Elastic net is a regularization method using both L1 and L2.

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i(\mathbf{x}_i^T\mathbf{w}+c))))+\lambda_1\|\mathbf{w}\|_1+\lambda_2\|\mathbf{w}\|_2$$

where $\|\mathbf{w}\|_2 = \sum_j (w_j)^2$

3) Sparse Group Regularization: Sparse Group regularization [12] penalizes model parameter $\mathbf{w}$ with the predefined structure of features formed as groups with L2 regularization in addition to L1 regularization for each feature. So parameters in same feature set are penalized together with the same amount of penalty.

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i(\mathbf{x}_i^T\mathbf{w}+c))) + \lambda_1\|\mathbf{w}\|_1+\lambda_2\sum_{j=1}^{g} b_j\|\mathbf{w}_{G_j}\|_2$$

$\mathbf{w}_{G_j}$ denotes non-overlapping groups of features. b_j denotes weight for group j.

IV. Experiment and Result

We experimented first with predict mortality on different feature sets and then we observed dynamics of topic proportions made from a topic feature.

topic 1: hematology	topic 2: skin	topic 3: respiratory(1)	topic 4: respiratory(2)	topic 5: cardiology	topic 6: chronic respiratory	topic 7: post-surgery	topic 8: hypotension	topic 9: heart-shock	topic 10: kidney
hct	skin	trach	cont	lasix	remain	pain	map	afib	assessment
bleed	wound	suction	remain	gtt	resp	drain	levoph	pain	response
unit	chang	thick	resp	heparin	vent	output	gtt	start	action
blood	area	vent	vent	ptt	cont	abdomin	renal	increas	failur
transfus	care	secret	chang	ccu	cchr	abd	crrt	rate	acut
receiv	impair	wean	amt	monitor	chang	fluid	neo	episod	continu
inr	dress	care	support	cath	coars	drainag	fluid	remain	renal
monitor	coccyx	psv	thick	chf	peep	monitor	wean	lopressor	monitor
cont	turn	respiratori	see	diuresi	will	abdomen	mcgkgmin	wean	respiratori
liver	continu	toler	abg	heart	sedat	prn	remov	bolus	cont

topic 11: nursing care(1)	topic 12: admission/transfer	topic 13: respiratory(3)	topic 14: nursing care(2)	topic 15: neurology	topic 16: nursing care(3)	topic 17: palliative care	topic 18: end-of-care plan	topic 19: general(1)	topic 20: general(2)
resp	hospit	sat	note	neuro	continu	increas	famili	sedat	continu
neuro	transfer	cough	tube	eye	care	continu	care	vent	temp
foley	known	mask	drain	movement	skin	team	comfort	abg	cultur
stool	admit	neb	right	head	note	awar	meet	wean	fever
clear	day	receiv	left	open	goal	stitl	support	fentanyl	monitor
soft	intub	micu	puls	pupil	remain	insulin	discuss	sat	sent
cont	found	lung	site	note	cchr	decreas	will	increas	tylenol
abd	now	floor	drainag	sbp	monitor	drop	remain	peep	blood
command	admiss	place	chest	gag	edema	improv	stitl	lung	vanco
nurs	treat	orient	line	left	area	blood	pts	fio	wbc

Fig. 3. Top 10 words for each topics within topic dynamics

A. *Predicting Mortality on Different Feature Sets*

With regard to regularization and classifier, we used [13]'s Sparse Learning with Efficient Projections (SLEP) library for implementation of the classifier with each regularization scheme. Accelerated gradient descent is used with backtracking line search to find optimal step size.

As you can see in the figure 1, the baseline of word feature present lowest classification power than any other feature sets except the concept feature. When looking at the results with no regularization first, the concept feature itself showed much lower predictability than word feature, but when it is coupled with word feature, the word+concept feature set shows improvement than word feature or concept feature itself. We believe that concept feature provided some information that word feature that might not have. When no regularization is applied, the topic feature shows the best performance. We can conjecture that probably LDA's performance on text data also shown in our dataset and classification problem. But when the regularization has applied the word + topic+ concept feature outperforms all other feature sets. This is probably because the additional information which might improve the classification is obtained when the original word feature is transformed into ontological feature and topic modeling is applied. In addition, regularization methods helped the model to find a linear decision boundary with less dimensional space reduced by the regularization methods. Note that all of our regularization settings has the lasso (L1) within its component which shuts off some features and results in the sparse representation of the original sample. However, when we compared different regularization methods, it is hard to find the much differences. Especially we can see that any regularization methods does not change the ranking of the AUC.

1) Detail on comparison of feature sets: Let's see more in detail on a comparison of feature sets. When the two derived feature sets, the concept feature and the topic feature, are compared, the concept feature underperforms than the topic feature. (When no regularization applied, AUC 0.83 of the topic feature and AUC 0.64 of concept feature shows this.) We assume that the lower predictability of the concept feature is caused by lower coverage of concept feature on certain samples. When a concept is mapped from the original text, only matched ones are included and any sample that does

not contain matched term perhaps causes loss of information. In contrast, LDA puts a small probability to samples even when it is not quite relevant to a topic. When we compare the word+concept feature and the word+topic feature without regularization, the word+concept feature shows slightly better performance (AUC 0.8416 of the word+concept and AUC 0.8378 of the word+topic feature). Considering lengthy training cost of LDA (around 6-7 hours with this size of the corpus), we perceive the merit of using additional information from the ontology. But when with regularization, the word+topic feature again outperforms the word+concept feature. One of the possible cause may be coming from the difference between topic and concept feature. On the other hand, one of the interesting aspects is that when all derived feature sets are combined together as the word + topic + concept feature set, it outperforms all other features. This shows the best AUC (0.8927) when it processed with the sparse group regularization with topic group weight adjusted +2.0. Also, the best PR-AUC (0.8103) is achieved with this feature set regularized with the Elastic Net.

2) Effect of Regularization: For the mixed features, the effect of regularization clearly showed at those features that contain the word features. It seems mainly because the L1 term of regularization eliminates inappropriate features and the logistic regression classifier can make a decision space in compact feature space. Unlike expectation, sparse group lasso seems does not make a large improvement in the performance of the classifier. But we can observe that changing group specific weight make a little difference in performance and it improves the topic+concept feature's AUC from 0.8632 with L1 to 0.8704 with the sparse group regularization.

B. Dynamics of Topic Proportions on Last 10 Days

In order to see the how the topic feature contributed to the predictability of patient mortality, we create another set of features. First, we aligned each patient's notes on the day of death. Then from the day of death, 48-hour time window moves backward and we merge any notes within the time window. The notes were irregularly written and there exist time gaps between notes even after the merging the process. Thereby, when a time gap between consecutive notes is more than 10 days, then the next note is disregarded. Subsequently, we retained patients only those who have more than 6 data points since less than this might be not enough to demonstrate changes of topics over time. This step results in 947 patients and 5682 notes. For the text preprocessing, we exclude word less than 20 occurrences in whole corpus and also set following terms as stopword: 'name', 'note', 'last', 'given', 'follow', 'patient', 'place', 'today', 'per', 'status','time', 'plan' It results in 5557 words in 5682 notes with 4,455,854 nonzero entries in word-note matrix.] Then, to obtain a lower-dimensional representation of clinical note, we use LDA and created topics. The number of topics is set as 20 for interpretability and each topic is labeled by a critical care physician afterward.

In figure 2, we can explicitly observe an upsurge of the topic on end-of-care plan (topic 18) from day 2-3 to the day of death. Also, the topic of palliative care (topic 17) smoothly increases as the time point close to the day of death. These points conform with the result of correlation test in 4. The topic of end of care plan and palliative care marked highest ranking on the correlation. On the other hand, topics related to post-surgery (topic 7), hematology (topic 1), cardiology (topic 5), nursing care 2 (topic 14) and topics on respiratory (topic 1, 2, 6) decreases as the day closes to the day of death.

Topic	Correlation with day close to death
topic18: end-of-care plan	0.8249
topic17: palliative care	0.8208
topic08: hypotension	0.7219
topic10: kidney	0.2786
topic02: skin	0.1041
topic16: nursing care(3)	-0.1294
topic03: respiratory(1)	-0.3301
topic06: chronic respiratory	-0.4609
topic09: heart-shock	-0.6002
topic13: respiratory(3)	-0.6283
topic11: nursing care(1)	-0.6633
topic19: general(1)	-0.7214
topic12: admission/transfer	-0.7708
topic15: neurology	-0.7806
topic04: respiratory(2)	-0.7966
topic20: general(2)	-0.8995
topic14: nursing care(2)	-0.9046
topic05: cardiology	-0.9357
topic01: hematology	-0.9448
topic07: post-surgery	-0.9653

Fig. 4. Correlation between topic proportions and day close to the death

V. CONCLUSION

In conclusion, we have observed that features obtained from medical ontologies and topic modeling improved the modeling of patient mortality using the clinical note. This shows us the combination of semantically enriched concepts from ontology and probabilistically dimensionality reduction methods has potential to enhance tasks that have done before separately. We also observed that regularizations have an effect on improving classification especially when the dimensionality of the feature is high by making the representation of feature sparser.

REFERENCES

[1] M. Ghassemi, et.al Topic Models for Mortality Modeling in Intensive Care Units. ICML Workshop in Clinical Data Analysis 2012.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022. 2013.

Initial Feature Sets

Feature set	Regularization	Feature size	Nonzero feature size	AUC	AUC Variance	PR-AUC	PR-AUC Variance
Word	(1) No Regularization	10229	10229	0.8304	0.0002	0.6670	0.0007
Word	(2) L1	10229	353	0.8561	0.0001	0.7257	0.0002
Word	(3) ElasticNet (L1+L2)	10229	307	0.8561	0.0001	0.7292	0.0003
Concept	(1) No Regularization	118	118	0.6488	0.0002	0.5200	0.0003
Concept	(2) L1	118	52	0.6557	0.0002	0.5325	0.0002
Concept	(3) ElasticNet (L1+L2)	118	50	0.6547	0.0001	0.5322	0.0002
Topic	(1) No Regularization	50	50	0.8720	0.0000	0.7854	0.0001
Topic	(2) L1	50	45	0.8725	0.0000	0.7860	0.0002
Topic	(3) ElasticNet (L1+L2)	50	43	0.8727	0.0000	0.7863	0.0002

Mixed Feature Sets

Feature set	Regularization	Feature size	Nonzero feature size	AUC	AUC Variance	PR-AUC	PR-AUC Variance
Topic + Concept	(1) No Regularization	168	168	0.8703	0.0001	0.7703	0.0002
Topic + Concept	(2) L1	168	63	0.8766	0.0000	0.7866	0.0002
Topic + Concept	(3) ElasticNet (L1+L2)	168	60	0.8758	0.0000	0.7801	0.0001
Topic + Concept	(4) Sparse Group Regularization	168	63	0.8766	0.0000	0.7866	0.0002
Topic + Concept	(4) Sparse Group Regularization (2nd group -0.5)	168	64	0.8776	0.0000	0.7878	0.0002
Topic + Concept	(4) Sparse Group Regularization (2nd group +2.0)	168	63	0.8752	0.0000	0.7818	0.0001
Word + Concept	(1) No Regularization	10318	10317	0.8416	0.0003	0.6813	0.0007
Word + Concept	(2) L1	10318	375	0.8632	0.0001	0.7404	0.0003
Word + Concept	(3) ElasticNet (L1+L2)	10318	328	0.8631	0.0001	0.7430	0.0003
Word + Concept	(4) Sparse Group Regularization	10318	375	0.8632	0.0001	0.7404	0.0003
Word + Concept	(4) Sparse Group Regularization (2nd group -0.5)	10318	602	0.8704	0.0001	0.7595	0.0001
Word + Concept	(4) Sparse Group Regularization (2nd group +2.0)	10318	1291	0.8669	0.0000	0.7627	0.0001
Word + Topic	(1) No Regularization	10279	10279	0.8378	0.0002	0.6779	0.0008
Word + Topic	(2) L1	10279	255	0.8873	0.0000	0.8011	0.0003
Word + Topic	(3) ElasticNet (L1+L2)	10279	206	0.8872	0.0000	0.8032	0.0004
Word + Topic	(4) Sparse Group Regularization	10279	255	0.8873	0.0000	0.8011	0.0003
Word + Topic	(4) Sparse Group Regularization (2nd group -0.5)	10279	461	0.8859	0.0000	0.8026	0.0003
Word + Topic	(4) Sparse Group Regularization (2nd group +2.0)	10279	6889	0.8875	0.0000	0.8066	0.0004
Word + Topic + Concept	(1) No Regularization	10397	10397	0.8498	0.0002	0.6936	0.0006
Word + Topic + Concept	(2) L1	10397	278	0.8925	0.0000	0.8076	0.0002
Word + Topic + Concept	(3) ElasticNet (L1+L2)	10397	222	0.8919	0.0000	0.8103	0.0003
Word + Topic + Concept	(4) Sparse Group Regularization	10397	278	0.8925	0.0000	0.8076	0.0002
Word + Topic + Concept	(4) Sparse Group Regularization (2nd group -0.5)	10397	491	0.8916	0.0000	0.8090	0.0002
Word + Topic + Concept	(4) Sparse Group Regularization (2nd group +2.0)	10397	285	0.8927	0.0000	0.8091	0.0002

Fig. 5. Experiment result on classifying patient mortality with different features and regularizations

[3] C. Howes, M. Purver, R. McCabe, Investigating topic modeling for therapy dialogue analysis. IWCS Workshop in Computer Semantic Clinical Text 2013

[4] P. Resnik, et al. Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. Computational Linguistics and Clinical Psychology Workshop 2015

[5] Y. Halpern, et al. A comparison of dimensionality reduction techniques for unstructured clinical text. ICML Workshop in Clinical Data Analysis 2012

[6] M. Ghassemi, et al. Unfolding physiological state: mortality modelling in intensive care units. KDD 2014

[7] A. Masino, T. Dechene, M. Dulik, A. Wilkens, N. Spinner, I. Krantz, J. Pennington, P. Robinson, and P. White, Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. BMC bioinformatics, 15(1), p.1. 2014 Vancouver

[8] S. Lu, Y. Ye, R. Tsui, H. Su, R. Rexit, S. Wesaratchakit, X. Liu, and R. Hwa, Domain ontology-based feature reduction for high dimensional drug data and its application to 30-day heart failure readmission prediction. In Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on (pp. 478-484). IEEE. 2013.

[9] C. Gillies, M. Siadat, N. Patel, and G. Wilson, A simulation to analyze feature selection methods utilizing gene ontology for gene expression classification. Journal of biomedical informatics, 46(6), pp.1044-1059. 2013.

[10] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark,MIMIC-III, a freely accessible critical care database. Scientific data, 3, 2016.

[11] A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium. 2001:17-21.

[12] J. Liu, and J. Ye. Moreau-Yosida Regularization for Grouped Tree Structure Learning. NIPS 2010

[13] J. Liu, S. Ji, and J. Ye. SLEP: Sparse Learning with Efficient Projections. Arizona State University, 2009. http://www.public.asu.edu/ jye02/ Software/SLEP.