External Evaluation of Topic Models: A Graph Mining Approach

Hau Chan
Department of Computer Science
Stony Brook University
hauchan@cs.stonybrook.edu

Leman Akoglu
Department of Computer Science
Stony Brook University
leman@cs.stonybrook.edu

***Abstract*—Given a topic and its top-k most relevant words generated by a topic model, how can we tell whether it is a low-quality or a high-quality topic? Topic models provide a low-dimensional representation of large document corpora, and drive many important applications such as summarization, document segmentation, word-sense disambiguation, etc. Evaluation of topic models is an important issue; since low-quality topics potentially degrade the performance of these applications. In this paper, we develop a graph mining and machine learning approach for the external evaluation of topic models. Based on the graph-centric features we extract from the projection of topic words on the Wikipedia page-links graph, we learn models that can predict the human-perceived quality of topics (based on human judgments), and classify them as high or low quality. Experiments on four real-world corpora show that our approach boosts the prediction performance up to 30% over three baselines of various complexities, and demonstrate the generality of our method to diverse domains. Further, we provide interpretation of our models and outline the discriminating characteristics of topic quality.**

***Keywords*-topic models; human evaluation; graph mining**

I. Introduction

Topic modeling is an area that focuses on the extraction of topics from document corpora. Given a large collection of documents D and the number of desired topics T, a topic modeling method M, such as LDA [1], models each document $d \in D$ as a multinomial distribution over T topics, where each topic is in turn a multinomial distribution over W words. Typically, only a small number of words are important (i.e. have high likelihood) in each topic (also only a small number of topics are relevant for each document).

Topic models have been studied widely [1], [2], [3] and have important applications in database summarization [4], word-sense discrimination [5], information discovery [6], and many others. Naturally these applications rely on the quality of topics that the topic models generate. An issue of concern, however, is that it is often likely for topic models to output low-quality topics in addition to the high-quality ones. For example, consider the two topics with their top 10 most likely words in Table I. From humans' perspective, the first topic (T1) consists of more semantically coherent words, while the second topic (T2) contains patchy groups of mostly incoherent words.

Low-quality topics can potentially degrade the performance of the applications; e.g. they could mislead topic-based document similarity, introduce noise in clustering, and cause poor semantic interpretation. This makes the evaluation of topic models a crucial task.

Previous research focused on the statistical (or quantitative) evaluation of topic models [7]. This type of evaluation measures either the generalization performance of a topic model based on likelihood on held-out test datasets, or its performance on external tasks. However these do not measure the interpretability of individual topics. In fact, the seminal paper [8] showed that there is a negative correlation between human evaluation and statistical evaluation of topic models. This finding started a new episode in topic model evaluation, by shifting focus to semantic coherence of topics. It prompted researchers to come up with conceptual (or qualitative) evaluation techniques, that can identify the human-perceived quality of topics.

Several works on conceptual evaluation of topic models have been proposed within the last 4-5 years [8], [9], [10], [11], which consider the coherence of individual topics. [8] elicits human input, while others try to develop a *single* statistical measure that mimics real human judgements on topic-evaluation tasks. (See §V for details on related work). None of these proposals (i) exploits a *collection* of evidential measures, or (ii) builds a learning model to predict conceptual topic quality; which is the basis of our work. We summarize our contributions below.

- *Problem formulation:* We formulate the evaluation problem as a supervised classification task and derive a predictive model that "learns" from human judgments to classify topics as good or poor as perceived by humans.
- *Novel graph-centric features using Wikipedia:* To construct a set of evidential features for our learner, we develop a novel graph mining approach which revolves around the creation and extraction of graph-centric properties of the topic-words' projection subgraphs on the Wikipedia page-links graph (referred to as *WikiLinks* throughout text). *WikiLinks* consists of nodes that represent "things" that have a Wikipedia page where edges capture the hyperlinks among these pages. Intuitively, we think of semantically coherent topics

Table I
EXAMPLE TOPICS T1 (HIGH-QUALITY) AND T2 (LOW-QUALITY) OF A TOPIC MODEL.

T1:	steam, engine, valve, piston, cylinder, pressure, boiler, air, pump, pipe
T2:	cut, system, capital, pointed, opening, building, character, round, france, paris

to consist of words that are "close-by" in this graph, and construct features based on graph topology and closeness accordingly.

- *Experiments:* Using our predictive model we perform experiments on topics extracted from four real-world document corpora: two from news, one from books, and one from medicine. Our results show the effectiveness and generality of our approach in predicting and interpreting the human-perceived quality of topic models, where we achieve up to 30% better classification performance compared to three baseline predictors.

In the rest of the paper, we give an overview of our proposed method (§II), explain it in detail (§III) present experiment results (§IV), survey related work (§V), and conclude with summary and future directions (§VI).

II. OVERVIEW

Problem Statement. The main research question we consider can be stated as follows:

Given a set of topics output by a topic model, how can we learn to classify them into low- versus high-quality topics (as perceived by humans)?

We give a more detailed definition of our problem in (§III-A) and provide the highlights of our proposed method next.

Proposed Topic Evaluation Framework. We describe our framework in five parts; (§III-B) Wikipedia page-links graph, (§III-C) graph projections, (§III-D) graph-centric features, (§III-E) labeled case libraries, and (§III-F) prediction models. A flow-diagram showing the operation of our method is given in Figure 1.

(§III-B) WikiLinks graph: Wikipedia is a large knowledge base being used and edited by millions of people around the world. To evaluate the human-perceived quality of topics, we use this knowledge base generated by humans themselves. Particularly we use the *WikiLinks* graph, in which nodes represent Wikipedia pages and edges denote their hyperlink relations.

(§III-C) Graph projections: Given a set of k topic words, we project the words onto the *WikiLinks* graph, that is, we *map* each topic word to the page that is *associated* with it in the graph. For example, the word `steam` in T1 above would map to page (or *WikiLinks* node) http://en.wikipedia.org/wiki/Steam. We then consider the induced subgraph of these nodes (projection graph), which may not be a connected subgraph. We choose several connector nodes in the original graph to obtain a second subgraph (spanning graph).

(§III-D) Graph-centric features: Wikipedia links graph is constructed by humans where editors introduce edges between pages (i.e. nodes in *WikiLinks*) by their relevance. Therefore, we expect the words of a semantically coherent topic to lie "close-by" in this graph, and extract features based on graph closeness.

(§III-E) Generating labeled case libraries: For our supervised classification task, we obtain binary training labels for topics (good versus poor quality). We consider two learning settings; one of predicting the relative quality of topics (based on rank order of words) where the labels are obtained implicitly, and another of predicting the absolute (i.e. human-perceived) quality of topics (based on human judgments) where labels are obtained explicitly.

(§III-F) Learning to predict topic quality: Finally, we use the graph-centric properties from (§III-D) as evidential features and the case libraries from (§III-E) as labels to learn statistical models that provide predictors of topic quality. Given many possible graph-centric features, we perform feature selection to identify a subset of discriminative ones, which we use to interpret our models.

III. PROPOSED FRAMEWORK

A. Problem Definitions

We consider the topic quality prediction problem. We study it under two settings: (1) absolute and (2) relative quality prediction.

Our main problem aims to build models to predict the absolute, or the human-perceived, quality of the topics. Here, scores provided by several human judges determine the positive and negative class training labels.

- **(P1) Absolute (Human-Perceived) Quality Prediction:** Given a topic (i.e. a set of k words), predict its quality (good/poor) as judged by humans.

We also study a related classification task of predicting relative quality of topic words. Each topic output by a topic model consists of a sequence of top-K words sorted by their relevance to the topic. In other words, the words ranked higher are more strongly related to a given topic than the words that come later in the sequence. We treat the top-k words of the topics as the positive (i.e. good) class examples, and bottom-k words in top-K as the negative (i.e. poor) class examples. For example for $k = 10$, bottom-k consists of words in rank order [11-20] when $K = 20$, and [91-100] when $K = 100$. As such, the prediction task becomes easier when K gets larger, since the separation between the training examples increases. Obtaining good prediction accuracy on

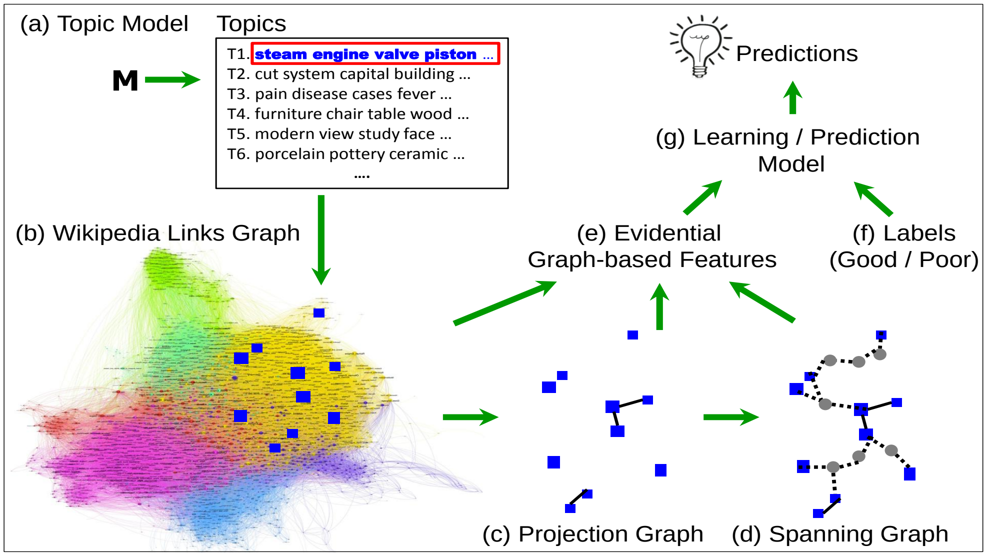

(a) Topic Model
Topics
M
T1. steam engine valve piston ...
T2. cut system capital building ...
T3. pain disease cases fever ...
T4. furniture chair table wood ...
T5. modern view study face ...
T6. porcelain pottery ceramic ...
....
Predictions
(g) Learning / Prediction Model
(b) Wikipedia Links Graph
(e) Evidential Graph-based Features
(f) Labels (Good / Poor)
(c) Projection Graph
(d) Spanning Graph

Figure 1. Proposed topic evaluation framework. Given output topics by a topic model, projection and spanning graphs of topic words are created based on the *WikiLinks* structure and graph-centric features are extracted for learning predictive models.

this task would prove *WikiLinks* a suitable external resource to rely on.

- **(P2) Relative Quality Prediction:** Discriminate good versus poor quality topics defined by top-k versus bottom-k words, respectively.

Having described our classification problems, we next need to represent each topic with a set of features. Our key idea for feature extraction is to exploit Wikipedia, and use this human-generated resource to construct topic subgraphs, from which we derive evidential topic features.

B. Wikipedia Links Graph

Wikipedia page-links dataset contains internal links between Wikipedia articles (i.e. entities)[1]. As such, the page-links data lends itself for a graph representation (which we call the *WikiLinks* graph) in which nodes denote Wikipedia entities, and edges capture the internal link relations among the Wikipedia articles.

For example, let us consider the entity `steam`. The corresponding Wiki-page can be found at http://en.wikipedia.org/wiki/Steam. Other Wiki-pages can be reached from this page by following hyperlinks on this page, e.g., the page on `piston` (http://en.wikipedia.org/wiki/Piston) and `mist` (http://en.wikipedia.org/wiki/Mist) are among those other, related entities. As such, the nodes `piston` and `mist` are 1-hop away from `steam`, thus are its neighbors.

WikiLinks is an excellent resource to guide for human-perceived evaluation of topic qualities, exactly because it is created by humans themselves—the entities are linked by their relatedness, as perceived by human editors.

Our key insight is to exploit the "graph-closeness" of related entities in *WikiLinks* to quantify the semantic quality of topics. Intuitively, the words of a semantically coherent topic, such as {`steam engine valve piston ...`}, would have high proximity in the *WikiLinks*. In fact, the wiki-page for `engine` directly links to `steam`, and `steam` links to `engine` through `steam engine`. putting these two words 1-2 hops away. In the sample visualization[2] of *WikiLinks* in Figure 1(b), related entities are observed to cluster in the graph topology.

As for coverage, Wikipedia provides a comprehensive resource with a massive collection of entities. In our version of *WikiLinks*[1], the graph statistics are:

	$\lvert N\rvert$	Directed $\lvert E\rvert$	Undirected $\lvert E\rvert$
WikiLinks	$17,170,893$	$158,373,970$	$117,434,138$

C. Projection and Spanning Graphs

We next provide definitions for topic subgraphs. Consider the *WikiLinks* graph $G(N, E)$ with node set N, edge set E (we experimented with both directed and undirected *WikiLinks*). Let W denote the set of k topic words, i.e. $|W| = k$. We project the topic words onto *WikiLinks* by *mapping* each word to a node (or entity) in the graph. In general, not all words will exist in *WikiLinks*, that is, $|N \cap W| \leq k$. We denote the mapped word set as $M = N \cap W \subseteq W$.

- **Topic projection graph** is a subgraph $g_M(M, E_M)$ induced on G with node set M and edge set E_M : $\{(u, v) \in E, u \in M \wedge v \in M\}$.

This graph may potentially consist of multiple disconnected components. In order to obtain a connected graph, we use a set of additional, connector nodes $C \subseteq N$ to build a graph that *spans* the topic words.

[1] http://wiki.dbpedia.org/Downloads38#wikipedia-pagelinks

[2] Resource: http://www.flickr.com/photos/mbiddulph/6070900906/

- **Topic spanning graph** is a subgraph $g_S(M \cup C, E_S)$ with node set $U = M \cup C$ and edge set $E_S : \{(u,v) \in E, u \in U \wedge v \in U\}$.

Ideally, the spanning graph contains the minimal set C to make the projection graph connected. However, it is NP-hard to find the minimal set, by reduction from the Steiner tree problem [12]): given a set X of nodes, interconnect them by a subgraph of shortest cost, where cost is defined as the sum of the (weights) of edges in the resulting subgraph. Therefore, we use the Minimum Spanning Tree (MST) approximation of the Steiner tree problem.

To construct the spanning graph, we first compute the pairwise shortest paths among the mapped M nodes to build a graph g_{SP} with edge weights $w(u,v)$, where $w(u,v)$ denotes the shortest path length in G between nodes $u, v \in M$. For undirected *WikiLinks*, g_{SP} is a complete graph as all nodes have paths from one to another (i.e. *WikiLinks* is a weakly connected graph). For directed *WikiLinks*, g_{SP} may contain missing edges as not all nodes have a *directed* path to others (i.e. *WikiLinks* contains multiple strongly connected components). Next we find the MST of g_{SP}, and *expand* it to obtain the spanning graph. Expansion involves introducing the connector nodes along the shortest paths of the MST, where we denote the union of connector nodes by C. Note that the spanning graph may no longer be a tree but may contain loops due to the intersection of the connector node sets of the paths.

Following on our running example, we show the projection and spanning graphs for the topics T1 and T2 of §I in Figure 2.

D. From Topic Subgraphs to Graph-Centric Features

Given the projection and spanning subgraphs, we generate a set of evidential graph-centric features. There are many features one could extract from a given graph. We want features that could potentially help differentiate good topics from poor ones. Good-quality topic words are conjectured to lie "close-by" in *WikiLinks*, reachable with many short paths from one another. On the other hand, the words of a poor topic would be separated in the graph topology. We can observe that these insights hold for T1 (good) and T2 (poor) of §I in Figure 2. Specifically, T1 contains more words that exist in *WikiLinks* (i.e. words that map to *WikiLinks* nodes), consists of fewer connected components in its projection subgraph (i.e. more nodes with direct connection), requires fewer connector nodes to build its spanning graph, and so on. Using these observations, we construct features based on graph topology and closeness.

Table II gives the list of features we constructed. In total, we constructed 19 features capturing the key topological properties of the projection and spanning subgraphs, as well as the closeness measures of the topic words in the original *WikiLinks* graph.[3] We group our features into three:

- PROJ contains features of the projection graph g_M, such as the maximum node degree, the number of connected components in g_M, etc.;
- D-SPAN consists of topological features of the directed spanning graph g_S including its density, ratio of connector nodes to mapped nodes, etc.;
- D-SP consists of features capturing the pairwise reachability between the topic words (excluding the self-pairs), based on the directed shortest paths.

Next we describe how we construct labeled case libraries for training and how we learn classification models for topic quality prediction.

E. Generating Case Libraries

We used news articles, books, and medical documents as our corpora. Descriptions of the datasets are in Table III.

For the prediction of the human-perceived (absolute) quality of topics (P1), we used the BOOKS and NEWS corpora, as previously used in [10], [11].[4] They consist of $T = 120$ and $T = 117$ topics, respectively. We considered the topics to consist of their top-10 words. All 237 topics were presented to 9 human judges. The judges were given guidelines on how to judge the *goodness* of the topics, and decide to what extent the topics were coherent, interpretable, meaningful, and easy-to-label with a short subject heading. They were also shown examples of good and bad topics. Notice that the BOOKS and NEWS corpora come from domains that are quite general, and thus we do not require the judges to have expertise in a specific domain (e.g., medicine).

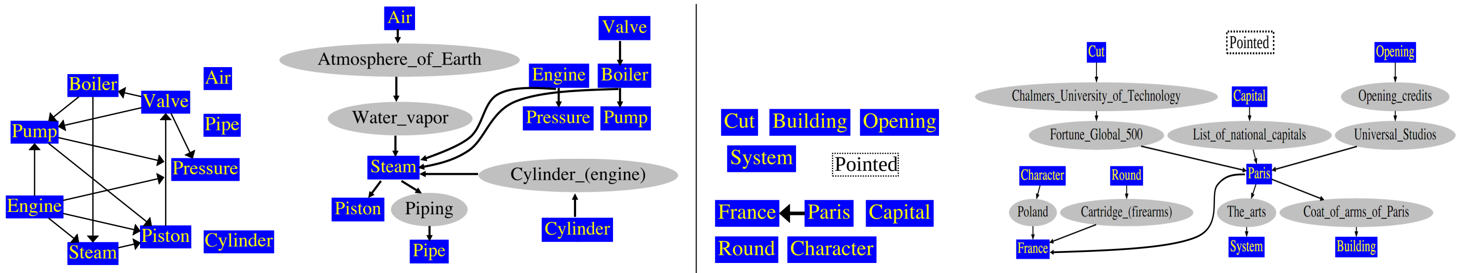
These nine judges evaluated the topics and provided annotations for each topic in 3-point scale: 1: 'good', 2: 'mediocre', 3: 'poor'. We used these human ratings to generate labels for our classification models. Specifically, topics with average rating below 1.5 are assigned to the positive (good) class, and negative otherwise. Examples of training topics from BOOKS (top few words) are given below (average rating in parentheses):

+	silk lace embroidery tapestry gold embroidered ... (1)
+	garden plant soil planting seed bloom spring ... (1.11)
+	seed trees soil root planting plant tree ... (1.33)
−	world people soul mind read reading live ... (2.56)
−	white munich phil room student people head ... (2.67)
−	person occasion purpose respect answer short ... (3)

We assume that there exists a global ground truth of labels and each human expert is a noisy version of it. To validate this, we performed two measurements to quantify the inter-annotator agreement among the nine judges. The average pairwise Spearman's rank correlation coefficient is

[3] Our experiments with directed and undirected versions of the *WikiLinks* graph revealed that the directed features provide more predictive power than the undirected ones. Therefore we focus our discussion on the directed features.

[4] We thank David Newman and his group for sharing the NEWS and BOOKS datasets as well as their human topic annotations.



(a1) g_M of T1 (b1) g_S of T1 (a2) g_M of T2 (b2) g_S of T2

Figure 2. Projection and spanning graphs, g_M and g_S respectively, for the two example topics T1 and T2 as given in §I. Blue square: mapped topic word, dotted white square: missing word, gray oval: connector node.

Table II
EVIDENTIAL FEATURES PROJ AND D-SPAN EXTRACTED RESPECTIVELY FROM PROJECTION AND SPANNING GRAPHS OF TOPIC WORDS ON *WikiLinks*, AS WELL AS PAIRWISE SHORTEST PATH FEATURES D-SP, ALL USED IN MODEL LEARNING. (MST: MINIMUM SPANNING TREE)

***WikiLinks* Feature**	**Description**
PROJ: Topic projection graph (g_M) features (4)	
$g_M NumMiss$	number of missing words in *WikiLinks* (i.e. $k - \|M\|$)
$g_M NumConnComp$	number of connected components in g_M
$g_M SizeMaxComp$	number of nodes in largest component of g_M
$g_M MaxDeg$	maximum node degree in g_M
D-SPAN: (Directed) Topic spanning graph (g_S) features (6)	
$g_S AvgMSTWeight$	average weight of MST (i.e. $W_{MST}/\|M\|$)
$g_S RatioC$	ratio of connector to original nodes in g_S (i.e. $\|C\|/\|M\|$)
$g_S MaxDegreeM$	maximum original node degree in g_S
$g_S MaxDegreeC$	maximum connector node degree in g_S
$g_S AvgDegree$	average degree of nodes in g_S
$g_S Density$	density of g_S (i.e. $\|E_S\|/(\|M \cup C\|(\|M \cup C\| - 1))$)
D-SP: (Directed) Shortest path features among topic word pairs (9)	
$NumNoPath$	number of pairs with no directed path inbetween
$AvgSPLen$	average pairwise directed shortest path length
$MaxSPLen$	maximum pairwise directed shortest path length
$NumSP1$	number of pairwise directed paths of length 1
$NumSP2$	number of pairwise directed paths of length 2
$NumSP3$	number of pairwise directed paths of length 3
$NumSP4$	number of pairwise directed paths of length 4
$NumSP5$	number of pairwise directed paths of length 5
$NumSP6+$	number of pairwise directed paths of length ≥ 6

found as $\rho = .73$ for NEWS and $\rho = .78$ for BOOKS. We also used Cohen's kappa, which provides a measure of the degree to which two judges concur in their respective sortings of items into mutually exclusive categories. As our categories are ordinal, where a human-rating of 1 is better than 2 which in turn is better than 3, we used a weighted version of the statistic. The average pairwise Cohen's kappa is found as $\kappa = .64$ for NEWS (max $\kappa = .79$), and $\kappa = .69$ for BOOKS (max $\kappa = .85$). Randomization tests yielded $\kappa = 0$ as expected. While there is no precise rule for interpreting kappa scores, [13] suggests that scores in the range $(.60, .80]$ correspond to "substantial agreement" between the annotators.

For the prediction of the relative quality of topics (P2), we used the publicly available PRESS[5] and BRAIN[6] corpora and learned topic models with $T = 100$ and $T = 200$ topics, respectively. We considered the top-10 words for each topic to be in the positive (good) class. For the negative class, we built three case libraries with words of ranks [11-20], [31-40], and [91-100]. This way we constructed three different learning tasks each with 200 and 400 training examples for PRESS and BRAIN, respectively. Examples of training topics from PRESS (top few words) are given below ([top 1-10] vs. [91-100]):

[5] http://www.cs.princeton.edu/~blei/lda-c/
[6] https://code.google.com/p/topic-modeling-tool/downloads/list

Table III

DATASETS USED IN OUR EXPERIMENTS. D: NUMBER OF DOCUMENTS IN THE CORPUS, T: NUMBER OF TOPICS, $Labels$: WHETHER HUMAN ANNOTATIONS EXIST OR NOT.

Dataset	D	T	$Labels$	**Description**
BOOKS	$12,000$	120	Yes	Books downloaded from the Internet Archive
NEWS	$55,000$	117	Yes	NYTimes news articles from LDC Gigaword
PRESS	$2,246$	100	No	Documents from the Associated Press
BRAIN	$10,000$	200	No	Pubmed abstracts for the query "brain injury"

+	space soviet shuttle nasa launch mission earth venus ...
−	jupiter day help report released days data laboratory ...
+	research scientists researchers animals project state ...
−	defense usda caused two temperatures side agricultural ...
+	power cars heat oil fuel energy electricity day ...
−	account total carbon year just united lower i plan ...

F. Learning to Predict

After constructing our topic libraries with labeled examples (§III-E) and extracting their graph-centric features (§III-D), we train logistic regression classifiers with L_1 norm regularization.

More specifically, we are given n training examples (n topics) $\{(x^{(i)}, y^{(i)}, i = 1, \ldots, n\}$, where each $x^{(i)} \in \mathbb{R}^m$ is an m dimensional feature vector, and $y^{(i)} \in \{1, 0\}$ denotes the class label (1: positive (good) vs. 0: negative (poor)). Logistic regression classifier models the probability distribution of the class label y given a feature vector x as $p(y = 0|x; w) = \sigma(w^T x) = \frac{1}{1+exp(-w^T x)}$, where $w \in \mathbb{R}^m$ are the parameters of the model (feature weights), and $\sigma(.)$ is the sigmoid function.

We regularize the logistic regression model using L_1 norm, which corresponds to Bayesian learning under the Laplace prior of the parameters; $p(w) = (\lambda/2)^m exp(-\lambda\|w\|_1)$, with $\lambda > 0$. The maximum a posteriori estimate of the parameters can be obtained by solving the (convex) optimization problem: $\arg\min_w \sum_{i=1}^n -\log p(y^{(i)}|x^{(i)}; w) + \lambda\|w\|_1$. As such, the Laplace prior "pushes" the weights towards zero, and biases the solution to be sparse. This helps us with feature selection. In order to solve for the model parameters w, and hence the feature "weights", we employ efficient algorithms [14] where we choose the hyperparameter λ using cross-validation. We report the leave-one-out cross-validation accuracies, which provide a good approximation to the true accuracy of our models. The results of our experiments are discussed in the next section.

IV. EXPERIMENT RESULTS

We start with discussing the performance results on the relative quality prediction problem. Later, we introduce two baseline techniques and proceed with human-perceived topic quality prediction results.

A. Relative Quality Prediction

As given in §III-A, the relative quality prediction problem (P2) deals with differentiating the top-ranked words of a given topic from its non-top-ranked words. The goal of this set of experiments is to understand the value of using the graph-centric evaluation framework we developed. Achieving promising performance on this pilot study would show us the feasibility of our approach.

In Table IV, we present the prediction accuracy of our model on the PRESS and BRAIN topics. The results are listed for our three different relative classification tasks (§III-E) and for our various groups of features (§III-D). From the tables, we observe that using our graph-centric features we achieve improved classification performance in all cases, and when features are used collectively we obtain 15% to 30% boost over the random baseline. As expected, the boost is gradually higher for the easier tasks (from left to right) where the negative class words are chosen further down in the rank order of topic words. These preliminary results show that *WikiLinks* is useful as an external resource and that our method is suitable for topic quality prediction tasks.

B. Building baselines

Before we move on to the results, we introduce two non-trivial baselines that we developed and compared to our approach, which are much smarter than the simple majority-class baseline.

1) Google baseline:: Given a set of k topic words, we used several Google operators[7] to query for results containing these words. In particular, from different types of Google queries, we built four what we call "Google features" per topic. Each Google feature is the logarithm of the number of webpages returned for its corresponding query, or in other words $\log(hitcount(query))$.

Table V gives a summary of the Google features we constructed. First, we queried for all the webpages that contain *all* the topic words in their text; by using the `allintext:`$word_1, word_2, \ldots, word_k$ operator. Second, we queried for the pages that contain *at least one* of the query words in their title; using `intitle:`$word_1$ `OR ... OR intitle:`$word_k$. Similar to the latter, we also queried for pages by their anchor

[7] http://www.googleguide.com/advanced_operators.html

Table IV
PRESS|BRAIN RELATIVE QUALITY PREDICTION RESULTS. CLASSIFICATION ACCURACIES FOR PREDICTING RELATIVE (TOP-k VERSUS NON TOP-k) TOPIC QUALITY, FOR VARIOUS GROUPS OF FEATURES.

Feature set top-10 vs.	top-[11-20]		top-[31-40]		top-[91-100]	
	PRESS	BRAIN	PRESS	BRAIN	PRESS	BRAIN
BASELINE-MAJORITY	0.500	0.500	0.500	0.500	0.500	0.500
PROJ	0.505	0.622	0.715	0.705	0.765	0.725
D-SPAN	0.650	0.687	0.760	0.740	0.805	0.762
D-SP	0.605	0.665	0.710	0.760	0.750	0.790
PROJ+D-SPAN	0.650	0.687	0.745	0.722	0.790	0.777
PROJ+D-SP	0.650	0.672	0.710	0.752	0.815	0.800
PROJ+D-SPAN+D-SP	0.660	0.687	0.735	0.752	0.810	0.807

or URL containment using inanchor:$word_1$ OR ... OR inanchor:$word_k$ as well as using inurl:$word_1$ OR ... OR inurl:$word_k$ to obtain the third and fourth features. As such, we represent each topic with four numerical features, and learn classification models based on those features.

Google features also rely on an external resource; the Google search engine. Unlike *WikiLinks* features, however, they do not exploit graph-centric properties of any projection or spanning graphs. We compare to this Google baseline to understand the amount of benefits gained by using the *WikiLinks* graph.

2) PPR baseline:: A second baseline classifier we built uses features based on the graph proximities among the topic words. To measure the proximity of a given pair of words on the *WikiLinks* graph, we used the personalized PageRank (PPR) scores [15]. Intuitively, the PPR score of a node v with respect to a given node u is high if there exist many, short paths between these two nodes. We constructed four PPR features capturing the pairwise graph-proximity between the topic words (excluding the self-pairs) as given in Table VI.

Table VI
PPR FEATURES GENERATED TO BUILD A BASELINE CLASSIFIER.

PPR Feature	**Description**
$AvgPPRscore$	average pairwise PPR score
$MedPPRscore$	median pairwise PPR score
$AvgPPRorder$	average pairwise PPR order
$MedPPRorder$	median pairwise PPR order

PPR-based features also exploit the underlying *WikiLinks* graph structure, and they are known as being more robust than shortest paths in capturing graph-centric proximities. As such, they build a strong baseline classifier. However, PPR computations are expensive as they rely on the mixing of random walks with restarts on the input graph (in tens of millions of nodes/edges). On the other hand, computing our graph features is fast since projected graphs are fairly small, and finding the shortest paths takes only a few seconds as often times the mapped nodes are close-by and thus most of the graph need not be traversed. Therefore we compare to PPR as a strong but expensive baseline, to understand its relative benefits compared to our method.

C. Absolute (Human-Perceived) Quality Prediction

As we motivated throughout the paper, the absolute quality prediction problem (P1 in §III-A) deals with differentiating the good quality topics from poor ones as perceived by human judges. We present our main results in Table VII.

We observe that *all* subsets of our feature groups outperform all three baselines. In particular, the Google baseline introduces 3-10% improvement in accuracy over the majority-class baseline, and the PPRbaseline based on the *WikiLinks* graph structure yields up to 23% increase. While these demonstrate the value of *WikiLinks* for this task, PPRbaseline is costly as we discussed earlier. On the other hand, all our graph-centric features introduce at least 25% and up to 30% boost over the majority baseline. In fact even the simplest group of our features PROJ, based on the immediate induced subgraph of topic words on the *WikiLinks*, outperforms the baselines alone.

We note that combined features do not always yield the best accuracy. We attribute this to the fact that learning with more features increases the size and complexity of our model space. With the same amount of data to learn from and a larger search space, our learning algorithm is less likely to find a good model, where having sufficiently large training data would mitigate this issue.

Cross-domain classification. In order to understand the generalization power of our framework, we also studied its cross-domain classification performance. Specifically, we learned a classification model using the BOOKS dataset and tested it on the NEWS dataset, similarly we also trained on NEWS and treated BOOKS as our test data. We show our results in Table VIII. The diagonal entries give the leave-out-out cross-validation accuracies within the same domain as before (last row in Table VII). The cross-domain accuracies are given on the off-diagonal entries. We observe that the cross-domain accuracies are fairly comparable to those of within-domain. This generalization power is partic-

Table V
Google features generated to build a baseline classification model.

Google Feature	**Description**
Operator:	Log-count of webpages that contain:
`allintext:`$word_1, word_2, \ldots, word_k$	all the topic words in their text
`intitle:`$word_1$ `OR ... OR intitle:`$word_k$	at least one topic word in their title
`inanchor:`$word_1$ `OR ... OR inanchor:`$word_k$	at least one topic word in their anchor
`inurl:`$word_1$ `OR ... OR inurl:`$word_k$	at least one topic word in their URL

Table VII
Books and News absolute quality prediction results. Accuracies for predicting absolute (human-perceived) topic quality, for various groups of features.

Feature set	Books	News	Books +News
Baseline-Majority	0.610	0.521	0.549
Baseline-Google	0.642	0.624	0.629
Baseline-PPR	0.842	0.735	0.785
PROJ	0.875	0.812	0.848
D-SPAN	0.892	0.769	0.844
D-SP	0.883	0.786	0.852
PROJ+D-SPAN	0.883	0.795	0.844
PROJ+D-SP	0.892	0.795	0.848
PROJ+D-SPAN+D-SP	0.900	0.821	0.831

Table VIII
Cross-domain absolute quality prediction results.

Test Train	Books	News
Books	0.900	0.769
News	0.867	0.821

ularly driven by our graph-centric features that are domain-independent.

Analysis of the prediction models. Finally, we study the characteristics of our learned models. As we use Lasso-regularization in our model training which lends itself to feature selection, we analyze the selected features (i.e. those with non-zero coefficients) for Books and News, as given in Table IX. We notice that the two models coincide in the majority of their selected features which hints towards the consistent evidential power of those features. One traditional way of interpreting the coefficients is to think in terms of the log-odds ratio, $\log \frac{P(y=0)}{P(y=1)} = w^T x$. Here, an increase of one unit in a particular feature i (under the same conditions for the others) contributes to the log-odds by w_i. Therefore features with positive coefficients contribute to the odds that a given topic is poor (i.e. $y = 0$), whereas features with negative coefficients advocate for the topic being good. More specifically, we deduce that good topics are those with fewer missing mapped words onto *WikiLinks* (or larger M), fewer connector nodes C in their spanning graphs g_S, and higher degree nodes in their projection graphs.

V. Related Work

Topic modeling has been a widely studied topic of interest especially for the machine learning (ML) [1], [16] (LDA, random projections), information retrieval (IR) [2], [3] (LSI, pLSA), as well as cognitive science [17] communities. Simply put, topic models describe the documents of a corpus as a mixture of topics, which in turn consist of a mixture of topic-words. Typically, only a small number of words are important in each topic, and only a small number of topics are present in each document. As such, topics provide low-dimensional representation for document collections [16] and drive many applications including document database summarization [4], segmentation [18], ontology learning [19], word-sense disambiguation [5], information discovery [6], to name a few.

Evaluation of topic models is an important issue, as unsupervised nature of the learning process makes model selection hard. Within the last 4-5 years[8], the natural language processing (NLP) community has shown increasing interest into the semantic coherence or in other words evaluation of topic models in capturing the human-perceived quality of topics. Accurately identifying and getting rid of low-quality topics not only would improve the understanding and interpretation of the semantic nature of topics, but it would also help boost the performance of many applications as listed above, e.g. better topic-based document similarity.

In this section we give a survey of topic model evaluation, and present related works in chronological order. Most works in quantitative evaluation of topic models [7] employ a variety of measures of model fit, such as estimating the likelihood of held-out documents or measuring the performance of an external task that is independent of the topic space such as information retrieval.

Drawbacks of model fit measures: While useful, these methods ignore the evaluation of the interpretability and semantic meaning of the topics for users. In fact, quite surprisingly, [8] showed that "traditional measures negatively correlated with the measures of topic quality" and that "models are often trading improved likelihood for lower interpretability".

Later, [10] proposed a new measure called pairwise mutual information (PMI) of topic-words based on co-occurrence statistics of word-pairs in large external text

[8] The first works on topic quality evaluation dates back to 2009 [8], [10].

Table IX

SELECTED FEATURES AND LEARNED COEFFICIENTS OF OUR L_1-REGULARIZED LOGISTIC REGRESSION MODEL FOR BOOKS AND NEWS. NEGATIVE (POSITIVE) COEFFICIENTS CONTRIBUTE TO THE ODDS OF A GIVEN TOPIC TO BE GOOD (POOR) QUALITY.

BOOKS		NEWS	
Selected Feature	**Coefficient**	**Selected Feature**	**Coefficient**
$g_M NumMiss$	0.0626	$g_M NumMiss$	0.0918
$g_S RatioC$	0.2940	$g_S RatioC$	0.5909
$g_M MaxDeg$	-0.2921	$g_M MaxDeg$	-0.4541
$g_M SizeMaxComp$	-0.8667	$g_S AvgMSTWeight$	0.2598
$NumSP2$	-0.9685		

corpora, and showed that PMI scores of topics are highly correlated (according to Pearson's correlation statistic) with human scores. [11] showed that PMI outperforms a range of other topic-scoring measures such as those based on lexical similarity and similarity in a given ontology. In [20], the PMI model is extensively evaluated on various different genres and domains of corpora (news, books, National Institutes of Health (NIH) abstracts) and various external corpora (Wikipedia articles, Google 5-grams, pubmed.gov abstracts).

Drawbacks of PMI-based evaluation: First, it requires the entire scan of external documents to compute the co-occurrence count for every pair of topic-words which is quite costly as the external corpora may be quite large (e.g., 2 million Wikipedia articles, 1 trillion Google 5-grams). Second, the best correlation to human-perceived quality depends on the type of external corpora used (according to [20], Google for books, Wikipedia for news, and pubmed.gov for NIH abstracts yield the best correlation). This makes it challenging to identify relevant, and burdensome to keep numerous corpora.

Rather than using external corpora, [9] proposed to use the original (i.e. training) corpus itself, which has been used for topic extraction, to compute a PMI-like score based on co-occurrence statistics of topic-words in the original document collection. Experiments on NIH document collection proved to be effective in separating low- and high-quality topics judged by domain experts. This is interesting, as the reasons behind *not* using the training corpus was stated in [10] as "...instead of using the collection itself to measure word association..., we use a large external text data source to provide *regularization*". This, of course, comes with the same challenges as for PMI. Recently [21] used cohesion and specificity of the topics to define a conceptual topic relevance score based on a concept hierarchy (i.e. an ontology).

Main drawback of existing methods: Relevance-based [21] and PMI-like measures [10], [9] as well as others compared to in [20] are all based on a *single* statistic. None of the methods exploit a *collection* of evidential measures to build a (learning) model that could potentially perform better than its parts. This is exactly the approach we take in this work.

Finally, while not directly applicable to evaluation, related work include automatic topic *labeling* [22], [23], [24] where the goal is to find a single most representative phrase (i.e. topic label or name) for each topic. Most related work in data mining that has inspired our work is [25], which used graph mining for evaluating the quality of search engine results to user queries. Other (although not directly) related graph-based techniques include connection subgraphs, with a goal of summarizing a subset of nodes [26], [27], [28].

VI. CONCLUSION AND RESEARCH DIRECTIONS

In this paper we introduced a novel graph mining approach for the external evaluation of topic models. We proposed to use Wikipedia as an external resource, constructed graph-centric features based on its page-links graph structure, and built classification models that can predict human-perceived quality of topics based on those evidential features. We summarize our contributions as follows.

- *Novel evaluation framework:* We develop a new topic quality evaluation framework that classifies a given topic as good or poor. It creates subgraphs of the topic words based on Wikipedia, and use their graph-centric properties to learn classification models.
- *Wikipedia as a knowledge base:* Wikipedia page-links graph consists of articles about entities, which are linked by their relatedness, as perceived by human editors. Thus we hypothesized that good, i.e. semantically coherent, topics' words would lie in close proximity in this graph. We validated this hypothesis with experiments that show significant improvements in prediction performance when *WikiLinks* is exploited.
- *Evidential graph-centric features:* We constructed novel features based on graph topology and closeness based on *WikiLinks*, in particular we introduced the projection and spanning subgraphs and their related features.
- *Prediction models and experiments:* Based on a carefully built list of features, we learned statistical classification models. Experiments highlighted the potential value of employing contextual subgraphs for understanding the quality of topics. One key aspect of our framework is its generality; it can be used with real-world corpora from diverse domains (e.g., news, books, and medicine), thanks to the immense coverage of Wikipedia as a knowledge base and domain-independent nature of our features.

Future work will look at graph-centric features describing the position of each mapped node in regard to other mapped nodes as well as to the rest of the graph, to identify specific topic words that are potentially out of context and outlying[9]. Another research direction is to exploit the *WikiLinks* graph structure to quantify the similarity of two or more topics by their positioning of words in the graph.

We believe that the presented work reveals an example where graph and data mining has potential impact to problems in related fields. We find that the proposed methods achieve desirable performance and provoke interesting directions for future research.

Acknowledgments

We thank the anonymous reviewers for their feedback. This material is based upon work supported by the Stony Brook University Office of the Vice President for Research and the National Science Foundation Graduate Research Fellowship. Any findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. of American Soc. for Info. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999, pp. 50–57.

[4] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization." in *NAACL*, 2009, pp. 362–370.

[5] S. Brody and M. Lapata, "Bayesian word sense induction." in *EACL*, 2009.

[6] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths, "Probabilistic author-topic models for information discovery." in *KDD*, 2004, pp. 306–315.

[7] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno, "Evaluation methods for topic models." in *ICML*, vol. 382. ACM, 2009, p. 139.

[8] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models." in *NIPS*, 2009, pp. 288–296.

[9] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models." in *EMNLP*, 2011, pp. 262–272.

[10] D. Newman, S. Karimi, and L. Cavedon, "External evaluation of topic models," in *Australasian Document Computing Symposium*, 2009, pp. 11–18.

[11] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *ACL*, 2010, pp. 100–108.

[12] R. M. Karp, "Reducibility among combinatorial problems." in *Complexity of Computer Computations*, 1972, pp. 85–103.

[13] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[14] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng, "Efficient L1 regularized logistic regression." in *AAAI*, 2006, pp. 401–408.

[15] T. H. Haveliwala, "Topic-sensitive pagerank," in *WWW*, 2002, pp. 517–526.

[16] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data." in *KDD*, 2001, pp. 245–250.

[17] T. L. Griffiths, M. Steyvers, and J. Tenenbaum, "Topics in semantic representation," *Psychological Review*, 2007.

[18] T. Brants, F. Chen, and I. Tsochantaridis, "Topic-based document segmentation with probabilistic latent semantic analysis," in *CIKM*, 2002, pp. 211–218.

[19] W. Wang, P. M. Barnaghi, and A. Bargiela, "Probabilistic topic models for learning terminological ontologies." *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 1028–1040, 2010.

[20] D. Newman, Y. Noh, E. M. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries." in *JCDL*, 2010, pp. 215–224.

[21] C. C. Musat, J. Velcin, S. Trausan-Matu, and M.-A. Rizoiu, "Improving topic evaluation using conceptual knowledge." in *IJCAI*, 2011, pp. 1866–1871.

[22] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models." in *KDD*, 2007, pp. 490–499.

[23] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models." in *ACL*, 2011, pp. 1536–1545.

[24] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia." in *WSDM*, 2013, pp. 465–474.

[25] J. Leskovec, S. T. Dumais, and E. Horvitz, "Web projections: learning from contextual subgraphs of the web." in *WWW*. ACM, 2007, pp. 471–480.

[26] L. Akoglu, J. Vreeken, H. Tong, D. H. Chau, N. Tatti, and C. Faloutsos, "Mining connection pathways for marked nodes in large graphs," in *SIAM SDM*, 2013.

[27] C. Faloutsos, K. S. McCurley, and A. Tomkins, "Fast discovery of connection subgraphs," in *KDD*, 2004, pp. 118–127.

[28] H. Tong and C. Faloutsos, "Center-piece subgraphs: problem definition and fast solutions," in *KDD*, 2006, pp. 404–413.

[9] Topic outlier words would be similar to the intrusive words in [8].