An Information Theoretic Approach to Quantifying Text Interestingness

Michal Dereziński
Computer Science Department
University of California, Santa Cruz
CA 95064, U.S.A.
mderezin@soe.ucsc.edu

Khashayar Rohanimanesh
eBay Research Labs
2145 Hamilton Avenue
San Jose, CA 95032, U.S.A.
krohanimanesh@ebay.com

Abstract

We study the problem of automatic prediction of text interestingness and present an information theoretic approach for quantifying it in terms of topic diversity. Our hypothesis is, in many text domains, often an interesting concept is generated by mixing a diverse set of topics. Given a word distributional model, we present an approach that leverages *Jensen-Shannon* divergence for measuring text diversity and demonstrate how such a measure correlates with text interestingness. We describe several different base-line algorithms and present results over two different data sets: a collection of e-commerce products from *eBay*, and a corpus of *NSF* proposals.

1 Introduction

With the rapid growth of e-commerce, new products are increasingly populated into the market place on daily basis. A larger subset of these products consists of our daily needs or off-the-shelf products, while a much smaller subset can be attributed as *unique*, *creative*, *serendipitous*, or *interesting* (see Figure 1). This class of products often provoke an emotive response in users and create a more engaging experience for the them (see *Pinterest* for example). Automatic discovery of this type of products is an important problem in e-commerce for creating an engaging experience for the users. Quantifying interestingness, however, is a challenging problem. There has been considerable research on visual interestingness and aesthetic quality of images [3, 9, 8, 5, 13, 18]. In text domain, researchers have studied different dimensions of this problem in terms of *humor identification* [12, 4, 10, 11], *text aesthetics* [14, 15, 7], and *document diversity* [1]. In this paper we only focus on text. Our hypothesis is, many interesting texts often present diversity in the text describing them. In examples shown in Figure 1, we have highlighted words that offer the largest diversity in each case. For example in Figure 1a, in the context of iPhone cases, one would expect less to observe topics that relate to makeup. In this paper we present an information-theoretic approach for measuring topic diversity based on *Jensen-Shannon Information Diversity* and show how it correlates with text interestingness. Measuring topic diversity in text has been previously studied by [1]. We show how our method differs from this approach and present empirical results over two different data sets: a collection of products from *eBay*, and a corpus of *NSF* proposals.

2 Our Approach

We assume a distributional representation for the words in the vocabulary (for a brief review see [17]). A distributional representation over a vocabulary V maps a word in the vocabulary to a probability distribution over a fixed set of contexts C. Often we start by a co-occurrence matrix $\mathcal{M}_{|V|\times|C|}$ where each row represents a co-occurrence of a word with the set of contexts C. For

(a) **Eyeshadow** Palettes for **iPhone** 6 case

(b) White Silicone **Horn** Stand Speaker for Apple **iPhone** 4/ 4S

(c) **Equation** Wall **Clock** Gifts for Math Gurus

Figure 1: A collection of unique/interesting *eBay* products. Highlighted keywords demonstrate how the text associated with such products could span multiple diverse topics.

example if we chose C=V then we obtain the familiar *word-to-word* co-occurrence representation which counts the number of times two words co-occurred in a document corpus. Another choice is to use the set of topics learned over a document corpus by *Latent Dirichlet Allocation* (LDA) [2] as the context. The rows of $\mathcal{M}_{|V|\times|C|}$ are then normalized in order to obtain a probability distribution over the context. We will use the notation P_w to represent the probability distribution over the context given a word w. We also use the notation $W = \{w_1, ..., w_k\}$ for a bag of word representation of a text snippet. Given a word distributional representation, $\mathcal{P}_W = \{P_{w_1}, ..., P_{w_k}\}$ denotes the set of probability distributions over the context for the words in the text snippet.

2.1 Information Diversity

Given a distributional representation over a vocabulary V and context C with P_w giving a probability distribution of a word w over the context C, we can measure information diversity for a text snippet $W = \{w_1, ..., w_k\}$ and its distributional representation $\mathcal{P}_W = \{P_{w_1}, ..., P_{w_k}\}$ as follows:

Definition 1 *Given a distribution P_w, its importance with respect to a prior distribution P is defined as $D_w = D_{KL}(P_w \| P)$ where $D_{KL}(\|)$ denotes the Kullback-Leibler Divergence.*

Definition 2 *Given a set of distributions $\mathcal{P}_W = \{P_{w_1}, ..., P_{w_k}\}$ and a prior P, we define a mixture distribution $P_W = \sum_{i=1}^{k} d_{w_i} P_{w_i}$ where $d_{w_i} = \frac{D_{w_i}}{\sum D_{w_j}}$ are the normalized importances.*

Essentially, P_W is the weighted average of the set $\mathcal{P}_W$, where the weights are chosen according to the importances. Next we define the diversity measure:

Definition 3 *We define the Jensen-Shannon Information Diversity of a set of distributions $\mathcal{P}_W$ with respect to prior P as $JSD_P(\mathcal{P}_W) = \sum_{i=1}^{k} d_{w_i} D_{KL}(P_{w_i} \| P_W)$ where d_{w_i} and P_W are as in the previous definition.*

This definition is closely related to the *general Jensen-Shannon Divergence* [6]. Another interesting theoretical property of Jensen-Shannon Information Diversity is that it can be interpreted as a generalization of Shannon entropy as a population diversity measure, however we will not go into this here any further.

2.2 Topic Diversity

In order to apply this model to natural language we first need to build a distributional representation for words. One natural choice for measuring the topic diversity is to use *word-to-topic* distribution. We need to address the following problems:
(1) Building word to topic distribution: We train an LDA [2] to build a topic model given a document corpus $\mathcal{C}$. From this LDA topic model we obtain the word-topic-count matrix $\mathcal{M}$ where $\mathcal{M}_{ij}$ is the number of assignments of j-th topic to word w_i in the corpus. By normalizing the rows of matrix $\mathcal{M}$ we will obtain a word-topic distribution, where the i-th row of the matrix gives the topic distribution for the word w_i (note that word-to-topic distribution as described above is not explicitly defined as a part of the standard LDA model and our approach is one way to approximate it). We

also use T to denote the set of topics learned in the LDA model.
(2) Obtaining a prior topic distribution: this is required for computing the information diversity as described in Section 2.1. We obtain a prior topic distribution P by computing the proportions of overall topic assignments. This corresponds to summing up matrix $\mathcal{M}$ along its rows and then normalizing the resulting vector.
(3) Capturing topic similarity: here we consider the problem first raised in [1]. When building a word-to-topic distribution model based on the word-to-topic co-occurrence matrix, the relationship among topics (i.e, topic similarity) may be lost. For example, if a word w has a topic distribution P_w concentrated on some topic t, then it will not necessarily peak at all other topics that are very similar to topic t. To address this problem, we first define a topic similarity matrix $\mathcal{S}_{|T|\times|T|}$ where the i^{th} row of $\mathcal{S}$ gives the topic similarity vector for the i^{th} topic (e.g., cosine similarity between topics [1]). We further assume that each row is normalized and hence can be thought of as a topic similarity distribution. Using $\mathcal{S}$ we obtain $\tilde{P}_w = P_w\mathcal{S}^T$ which essentially diffuses the initial distribution to one where all topics similar to some topic t are well represented. Similarly, we can use $\mathcal{S}$ to reflect the topic similarity in the prior distribution as $\tilde{P} = P\mathcal{S}^T$. In Section 3 we will show how this approach enhances the standard entropic measure of diversity.
(4) Sample size bias problem: note that when we normalize the rows of the matrix $\mathcal{M}$ to obtain a word-to-topic distributional model, the normalization factor for every word is simply the word count. Thus for a word w which occurs very rarely in the entire corpus $\mathcal{C}$, the topic distribution will be artificially skewed and is inaccurate simply because we do not have enough data points to estimate its true word-to-topic distribution. Now, if for this corpus it happens that the prior topic distribution is close to uniform, it will give a very high importance to the word w when measuring the information diversity as described in Section 2.1. One way to alleviate this problem is to use the relative sample size (e.g., word count) to smooth the distribution obtained by normalizing the rows of matrix $\mathcal{M}$. A natural choice for the smoothing distribution would in this case be simply the prior distribution mentioned above. Applying *Laplace smoothing* we get:

$$\widehat{P}_w = \frac{\alpha\tilde{P} + \mu_w\tilde{P}_w}{\alpha + \mu_w} \tag{1}$$

where μ_w is the frequency of word w in D, and P_w is the topic assignment distribution obtained from the word-topic matrix, while α is the parameter that specifies the strength of the prior.
(5) Conditioning on context words: we propose a final enhancement to the word-topic distributions. Suppose, the set of words $W = \{w_1, ..., w_k\}$ represents a text snippet that we want to analyze. The word w_i has a specific meaning inside of W, that can be significantly different than its meaning out of context. Denote $W_{\bar{i}} = W - \{w_i\}$ as the set of all words in W except w_i. By $P_{W_{\bar{i}}}$, we denote the mixture distribution for $W_{\bar{i}}$ (Definition 2) and we use $\tilde{P}_{W_{\bar{i}}}$ when it is smoothed using Laplace smoothing method. We propose the following definition of context-dependent word-topic distribution:

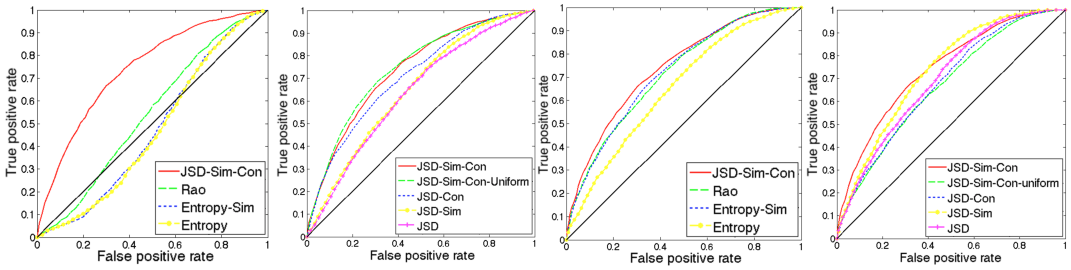
Definition 4 *Let $\tilde{P}, \widehat{P}_{w_i}, \tilde{P}_{W_{\bar{i}}}$ be the topic prior, general topic distribution for w_i, and the context distribution, respectively. Then, the context-dependent distribution is*

$$P_{w_i}^{W_{\bar{i}}}(t) \propto \frac{\tilde{P}_{W_{\bar{i}}}(t)}{\tilde{P}(t)}\widehat{P}_{w_i}(t)$$

There is a probabilistic explanation that we have left out for lack of space. However this can be intuitively understood as follows: we can think of $\frac{\tilde{P}_{W_{\bar{i}}}(t)}{\tilde{P}(t)}$ as a weight that further reshapes the smoothed word-to-topic distribution $\widehat{P}_{w_i}(t)$ to take into account the context. In our experiments we also smooth this distribution using *Laplacian* smoothing.

3 Experiments

We used the following two datasets in our experiments: (1) Interesting iPhone cases: for generating the ground truth data we hired workers from *Amazon Mechanical Turk (AMT)* to label a collection of nearly 20,000 iPhone cases on *eBay*. The details of this step is beyond the scope of this paper, however we used insights from interesting iPhone cases found on *Pinterest* and *eBay's* user behavior



(a) eBay (baseline) (b) eBay (JSD) (c) NSF (baseline) (d) NSF (JSD)

Figure 2: ROC curves presenting the results of experiments on the eBay dataset (a,b) and NSF proposal dataset (c,d). The comparison plots (a,c) show the results for our approach (JSD-Sim-Con) against other methods, while the plots (b,d) show different variations of our approach.

Table 1: Classification results for the eBay dataset.

	Precision	Recall	F1	Accuracy
JSD Features	$\mathbf{0.714} \pm 0.015$	0.597 ± 0.016	0.650 ± 0.014	$\mathbf{0.8828} \pm 0.0045$
RAE	0.676 ± 0.005	$\mathbf{0.666} \pm 0.030$	$\mathbf{0.671} \pm 0.013$	0.8809 ± 0.0020
SVD Features	0.676 ± 0.008	0.633 ± 0.017	0.654 ± 0.010	0.8778 ± 0.0027

data in order to generate a balanced data-set. We then pulled our final dataset from the annotated by selecting only those instances where the annotators all labeled it as positive (i.e., interesting) or negative (i.e., uninteresting). The final data-set consists of 2179 positive and 9770 negative instances for a total of 11,949 instances. For each instance, the product title of the corresponding *eBay* listing was used as the input. In this case we are dealing with very short text snippets, usually 10 to 12 words each. To train a topic model, we used a larger, more broader set of about 2 million product titles, grouped based on *eBay* categorical information into about 8,000 documents of approximately 200 titles each; (2) *NSF* abstracts: for the second dataset we used a set of 61,902 National Science Foundation Scholarship proposal abstracts (see [1] for more details) to evaluate how our diversity measure compares to other methods on larger pieces of text. We used this set for training a topic model, however to get labeled data, we had to generate artificial examples, by randomly mixing pairs of abstracts that we could expect to be either similar (small diversity) or very different (high diversity) and labeling them accordingly. We generated 5,000 of those examples with positive and negative labels evenly represented. For both datasets, we used the Mallet LDA implementation and learned a separate topic model with 400 topics.

We present two sets of results. First, we present ROC curves comparing different entropic measures of topic diversity in an unsupervised setting (labeled data is only used for generating the curves). Figures 2a and 2c compare our diversity metric using both topic similarity and context conditioning (labeled by *JSD-Sim-Con*) with a few baselines; namely LDA topic entropy, LDA topic entropy using topic similarity (labeled by *Entropy-Sim*), *Rao diversity* (see [1] for details). In either case it can be observed that our diversity metric outperforms the other baselines with an AUC around 0.73. Moreover, for the *eBay* dataset the other measures give poor results. This can be explained as follows: since the text snippets are short, the LDA may yield a poor topic inference for such short text and as a result all measures using topic inference would perform poorly. Figures 2b and 2d show the gains we obtain by applying topic similarity and context conditioning techniques (steps 4 and 5) that we discussed in Section 2.2. However, their degree of effects is different for each dataset. In the second set of results, we used the unnormalized vector of mixture topic distribution (described in Definition 2) computed over *eBay* product titles in a supervised classification setting. Table 1 shows the performance of the SVM classifier using our proposed mixture topic distribution as features and compares it to two different baselines, namely, SVM using *Latent Semantic Indexing (LSI)* features (by forming a document-term matrix and performing SVD), and a deep learning approach using the *recursive auto-encoders (RAE)* framework described in [16]. These results are averaged over five different cross-validation splits using 0.6 for training and 0.4 for testing. Our proposed approach shows a higher precision and a marginally higher accuracy compared to the baselines.

References

[1] Kevin Bache, David Newman, and Padhraic Smyth. Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 23–31, New York, NY, USA, 2013. ACM.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part III*, ECCV'06, pages 288–301, Berlin, Heidelberg, 2006. Springer-Verlag.

[4] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[5] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1657–1664, Washington, DC, USA, 2011. IEEE Computer Society.

[6] B. Fuglede and F. Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *IEEE International Symposium on Information Theory*, pages 31–31, 2004.

[7] Debasis Ganguly, Johannes Leveling, and Gareth Jones. Automatic prediction of aesthetics and interestingness of text passages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 905–916, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[8] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems*, 2011.

[9] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 419–426, Washington, DC, USA, 2006. IEEE Computer Society.

[10] Chloé Kiddon and Yuriy Brun. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL11)*, pages 89–94, Portland, OR, USA, June 2011. ACM ID: 2002756.

[11] Igor Labutov and Hod Lipson. Humor as circuits in semantic networks. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 150–155, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[12] Rada Mihalcea and Carlo Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 531–538, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[13] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 2049–2058. ACM, ACM, 2013.

[14] Jrgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE T. Autonomous Mental Development*, 2(3):230–247, 2010.

[15] Ekaterina Shutova and Lin Sun. Unsupervised metaphor identification using hierarchical graph factorization clustering. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 978–988. Association for Computational Linguistics, 2013.

[16] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[17] Joseph Turian, Dpartement Dinformatique Et, Recherche Oprationnelle (diro, Universit De Montral, Lev Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semisupervised learning. In *In ACL*, pages 384–394, 2010.

[18] Daphna Weinshall, Alon Zweig, Hynek Hermansky, Stefan Kombrink, Frank W. Ohl, Jrn Anemller, Jrg-Hendrik Bach, Luc J. Van Gool, Fabian Nater, Toms Pajdla, Michal Havlena, and Misha Pavel. Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. *IEEE Trans. Pattern Anal. Mach. Intell.*, (10):1886–1901.