Discovering Surprising Documents with Context-Aware Word Representations

Michał Dereziński
University of California
Santa Cruz, CA
mderezin@soe.ucsc.edu

Khashayar Rohanimanesh
eBay Research Labs
San Jose, CA
khash@cs.umass.edu

Aamer Hydrie
eBay Research Labs
San Jose, CA
ahydrie@gmail.com

ABSTRACT

User experiences can be made more engaging by incorporating surprise. For example, online shoppers may like to view unique products. In this paper we propose an approach for detecting surprising documents, such as product titles. As the concept of surprise is subjective, there is currently no principled method for measuring the surprisingness score of a document. We present such a method; an unsupervised approach for automatically discovering surprising documents in an unlabeled corpus. Our approach is based on a probabilistic model of surprise, and a construction of effective distributional word embeddings, which can be adapted to the semantic context in which the word appears. As the performance of our model does not degrade with the length of the document, it is particularly well suited for very short documents (even a single sentence). We evaluate our model both in supervised *and* unsupervised settings, demonstrating its state-of-the-art performance on two real-world data sets: a collection of e-commerce products from eBay, and a corpus of NSF proposals. These experiments show that our surprisingness score exhibits high correlation with human annotated labels.

ACM Classification Keywords

H.3.3 Information Search and Retrieval: Information filtering

Author Keywords

product discovery; text surprisingness; information theory; recommender systems; topic modeling

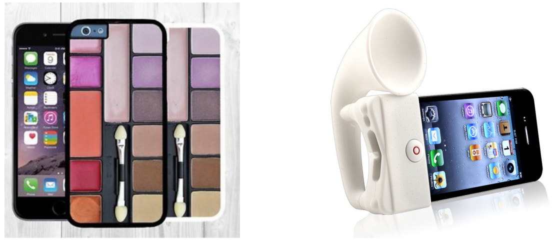
INTRODUCTION

Figure 1. Examples of off-the-shelf eBay products. Note that in every case, the word co-occurrences in the titles match our expectation.

(a) Black Qi Standard Wireless Charging Charger Receiver Case For iPhone 5 5G

(b) Bluetooth Wireless Speaker Mini Portable Super Bass For iPhone

Figure 2. Examples of surprising eBay products. The highlighted keywords are not expected to co-occur in the context of the product.



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

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

With the rapid growth of e-commerce, new products are increasingly introduced into the market place on a daily basis. A larger subset of these products consists of our daily needs or off-the-shelf products (Figure 1), while a much smaller subset can be attributed as *unique*, *creative*, *serendipitous*, or *surprising* (Figure 2). The latter class of products often provokes an emotive response in users and gives them a more engaging experience. Automatic discovery of this type of products is an important problem in e-commerce for creating an engaging experience for the users and encouraging them to return to the site. One approach to solving this task is through analyzing textual information that is associated with each product (e.g. title, description). Such a task can be framed as follows: given a large corpus (where each element describes one product), detect documents that are deemed surprising. This problem presents several challenges. First, the notion of surprise is not well understood and very subjective, two users may disagree about whether a product is surprising. Second, textual information is fairly limited in this domain. Many products may only be described by their title (typically 10 to 12 words), so the corpus may contain very short documents, making most text processing techniques ineffective. However, we believe that even from product titles alone it is possible to automatically detect items that appear unique and surprising to a sufficiently large population of users.

In this paper we present a probabilistic approach for discovering surprising documents. Unlike existing approaches, this method is shown to be effective even for very short text snippets. We hypothesize that many text forms that could be deemed surprising often express considerable variety in terms

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

of the text describing them. As an example, in Figure 2(a), in the context of iPhone cases, one would not expect to observe topics that relate to makeup. Based on this hypothesis we introduce a probabilistic model for measuring topic diversity based on *Jensen-Shannon Divergence* and show how it correlates with text surprisingness. To make our measure effective for short documents, we construct distributional word representations, which can adapt to the semantic context of the document. We present results evaluating the performance of our surprisingness metric in two different real world domains: (1) identifying unique eBay products based on the surprisingness of their titles; (2) idenitifying the most cross-disciplinary National Science Foundation Scholarship proposal abstracts (introduced in [1]). Moreover, we compare our model for constructing context-aware distributional representations to other standard text-embedding techniques, including recursive auto-encoders [17].

RELATED WORK

There has been considerable research on the topic of discovering unique, interesting or surprising documents. Researchers have studied different dimensions of this problem in terms of *humor identification* [12, 4, 10], *text aesthetics* [15, 16, 7], and *document diversity* [1, 8]. [12], studies a computational approach for humor recognition by utilizing a set of humor-specific stylistic features such as alliteration, antonymy, and adult slangs. Some of these features, e.g., antonymy, in a limited way capture some sort of text diversity as we do not normally expect antonyms co-occurring in standard text. [4], proposes a semi-supervised approach for identifying sarcastic sentences in Twitter and Amazon. Their approach consists of two stages; a semi supervised pattern acquisition is used for identifying sarcastic patterns, and a classifier uses such patterns as features in a classification task. Such word-level stand-alone interestingness features are related to the word-saliency factor that is discussed in Section 3. [7], studies automatic prediction of text interestingness by utilizing a broader set of features such as word length, repetitions, polarity, part-of-speech, semantic distances, and somewhat simple treatment of topic generality and diversity.

Measuring topic diversity for text has been previously studied by [1] which is the most relevant work to our approach. [1] uses Rao's diversity [13] for measuring document diversity based on a topic model learned over a corpus of documents. As shown in Section 5, we found that this method did not work well for the eBay product dataset. We believe that the primary reason for this is that the amount of useful textual data for each product is often too small to be treated as a document. As made clear by Sections 3 and 4 our measure differs from this metric from an information theoretic perspective, allowing it to perform well in this setting. Throughout experiments presented in Section 5 we use the approach of [1] as one of our baselines.

QUANTIFYING SURPRISE IN TEXT

The term "surprise" is not well defined and may vary depending on the domain. However, one concrete and general definition is offered through the probabilistic measure of "entropy" [14]. In the context of a text document, we can think of entropy as measuring the diversity of the topics/ideas which are being conveyed in the text. If we encounter two topics which are not expected to co-occur together, this surprises readers. Our goal is to define a quantifiable metric of entropic surprise in the text domain, which will perform well in practical tasks. Here, this problem is formulated as an unsupervised one, i.e. we expect our model to discover surprising documents in the absence of labeled training data or any user-specific information. We are given a large and representative corpus of documents from the considered domain, but no human-curated surprisingness labels are provided. We also require that the trained model be capable of efficiently scoring new documents as they come in, without the need to retrain on the entire corpus.

As our primary domain of interest is e-commerce, the documents we encounter are very short (e.g. product titles on eBay), often just 10 to 12 words. This makes standard bag-of-words models ineffective, because we cannot retrieve accurate word-frequency information for a single document. To address this problem, we can construct distributional word representations, which capture semantic information by embedding words into a multi-dimensional space of probability vectors.

Distributional Word Representations

As a starting point to obtaining distributional word representations, we run *Latent Dirichlet Allocation* (LDA) [2] to build a topic model from a corpus of documents. We use the set of topics as the contexts. In this case, each word is mapped to a probability distribution over the topics discovered by LDA. We will use T to denote the set of learned topics. As a result of this procedure, each word occurence in the corpus is tagged with one of the topics $t \in T$. From this model we obtain the word-topic-count matrix $\mathbf{X} \in \mathbb{R}^{|V|\times|T|}$ where entry x_{ij} is the number of assignments of topic t_j to word w_i from the vocabulary V. By normalizing the rows of matrix $\mathbf{X}$ we obtain a set of distributional representations, where the i-th row of the matrix gives a topic distribution for the word w_i.

Capturing topic correlations. When building a topic distribution model based on the word-to-topic co-occurrence matrix, the relationship among topics (i.e, topic correlation) may be lost. For example, a product description that is related to two topics such as *Computers* and *Software* should be deemed less surprising then if those two topics were *Computers* and *Clothes*, so our model has to be able to make such a distinction. To address this problem, we define a topic similarity matrix $\mathbf{S} \in \mathbb{R}^{|T|\times|T|}$ where the element $s_{i,j}$ gives the similarity/correlation between topic t_i and topic t_j (e.g. the similarity would be high between *Computers* and *Software*, but low for *Computers* and *Clothes*). Using $\mathbf{S}$ we transform the word-topic-count matrix $\mathbf{X}$, obtaining $\widehat{\mathbf{X}} = \mathbf{X}\,\mathbf{S}^\top$, which effectively diffuses the topic distributions to ones where for every topic assignment, all similar topics are also well represented. To construct the topic similarity matrix for the LDA model, we use the document-topic-count matrix $\mathbf{Y} \in \mathbb{R}^{|C|\times|T|}$ (where C is the set of documents), which is returned by LDA alongside $\mathbf{X}$. This matrix, for each document, has a row vector describing the total number of assignments of each topic. To estimate the similarity of topics t_i and t_j, we compute the cosine similarity between their corresponding columns in $\mathbf{Y}$. Denoting $\mathbf{y}_i$ and

$\mathbf{y}_j$ as those column vectors, their cosine similarity is

$$s_{i,j} = \frac{\langle \mathbf{y}_i, \mathbf{y}_j \rangle}{\| \mathbf{y}_i \|_2 \| \mathbf{y}_j \|_2},$$

where $\langle \cdot, \cdot \rangle$ represents a dot product between two vectors and $\| \cdot \|_2$ is the ℓ_2-norm of a vector.

Let $\widehat{\mathbf{x}}_w$ denote the row of matrix $\widehat{\mathbf{X}}$ corresponding to the word w. Applying *Laplace smoothing* to $\widehat{\mathbf{x}}_w$, we define the static representation of w:

$$P_w = \frac{\alpha P_0 + \widehat{\mathbf{x}}_w}{\alpha + \|\widehat{\mathbf{x}}_w\|_1}, \tag{1}$$

where $\| \cdot \|_1$ is the ℓ_1-norm of a vector, while α is the parameter that specifies the strength of the smoothing prior distribution P_0. We obtain a prior topic distribution P_0 by summing up matrix $\widehat{\mathbf{X}}$ along its columns and then normalizing.

Probabilistic Model of Surprise

For the remainder of this section, we assume that a topic distributional model for the words in the vocabulary has been learned. More specifically, we are given a distributional representation over a vocabulary V and topic set T, with P_w giving a topic distribution corresponding to the word w. We will measure surprise for a text snippet $W = (w_1, ..., w_k)$ by analyzing its distributional representation $\mathcal{P}_W = (P_{w_1}, ..., P_{w_k})$. To estimate the surprise induced by $\mathcal{P}_W$, we use a measure of divergence between probability distributions, discussed in [6], which is a variant of Jensen-Shannon Divergence.

Definition 1. *Let $\mathcal{P}_W = (P_{w_1}, ..., P_{w_k})$ be a sequence of distributions (for example, topic distributions for the words in a document). Moreover, assume $\{d_{w_i}\}$ is a set of importance weights assigned to $\{P_{w_i}\}$, such that $\sum_{i=1}^{k} d_{w_i} = 1$.* ***Jensen-Shannon Divergence*** *of the mixture $\Sigma d_{w_i} P_{w_i}$ is defined as*

$$D_{\mathrm{JS}}(\Sigma d_{w_i} P_{w_i}) = \sum_{i=1}^{k} d_{w_i} D_{\mathrm{KL}}(P_{w_i} || M),$$

where $M = \sum_{i=1}^{k} d_{w_i} P_{w_i}$ is the mixture distribution.

Here, D_{KL} corresponds to Kullback-Leibler Divergence (see [3]). As a shorthand, we will also write $D_{\mathrm{JS}}(M)$, when it is clear that M represents a mixture of distributions. It can be shown that Jensen-Shannon Divergence is a direct generalization of Shannon entropy.

Note, that to compute D_{JS} we need not only the distributional representations, but also a set of weights assigned to each of them. Intuitively, the weight d_{w_i} should describe the saliency of word w_i, i.e., how important it is in describing the content of the text (e.g. stop words should get a low weight). We want to quantify the information gain coming from each appearance of the word w_i. This can be described by the KL-divergence between the distributional representation of w_i and the prior topic distribution P_0.

Definition 2. *We define the* ***importance*** *of a word representation P_{w_i} with respect to the prior distribution P_0 as*

$$D_{w_i} = D_{\mathrm{KL}}(P_{w_i} \, || \, P_0).$$

For any text W we define its *static mixture representation* as $M_W = \Sigma d_{w_i} P_{w_i}$, where $d_{w_i} = D_{w_i} / (\sum_{j=1}^{k} D_{w_j})$.

CONTEXT-AWARE WORD REPRESENTATIONS

Let $W = (w_1, ..., w_k)$ represent 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. For example, the word *mouse* can be associated with the topic of computers as well as the topic of animals, which are clearly very different. However, in a given context, usually there will be no confusion as to which topic is most relevant. We propose to define a new context-aware distributional representation of w_i that takes this into account (as opposed to the static representation P_{w_i} we have so far). We will represent the context information using the static mixture distribution $M_{W_{-i}}$ (see Definition 2), where W_{-i} is the text W with word w_i excluded. If a topic t has high weight in both distributions P_{w_i} and $M_{W_{-i}}$, then it is relevant to w_i in this context. More generally, if distributions $M_{W_{-i}}$ and P_{w_i} have significant overlap, then those shared topics are likely to explain the text well, and the remaining topics in P_{w_i} should effectively be ignored. If the overlap is minimal, then this indicates unexpected (surprising) content. To model this intuition, we define the following context-aware representation of a word w_i in text W:

Definition 3. *Let w_i be a word in text W, and let P_0 be the prior topic distribution. Then the context-aware distributional representation of w_i in W is defined as*

$$\widehat{P}_{w_i}(t; W) \propto \left(\frac{M_{W_{-i}}(t)}{P_0(t)} + \beta \right) P_{w_i}(t).$$

This can be intuitively understood as follows: we can think of $\frac{M_{W_{-i}}(t)}{P_0(t)}$ as a weight that further reshapes $P_{w_i}(t)$ to take into account the context. The key operation here is the element-wise multiplication $M_{W_{-i}}(t) \cdot P_{w_i}(t)$, which essentially captures the overlapping topics. The prior P_0 is there to adjust for the overall popularity of each topic - if context distribution were equal to the prior, we would expect it to have no effect in this transformation, as is the case here. Note, that we apply additional Laplace smoothing with parameter β, to account for the case where the word and the context have minimal topic overlap, which is very important because that is what generates surprise. If this occurs, the transformation effectively reverts to the static representation.

Finally, we can define our surprise score by combining the context-aware word representations into a mixture using weights d_{w_i} (see Definition 2):

Definition 4. *We define the* ***text surprisingness*** *of W w.r.t. corpus C to be $D_{\mathrm{JS}}(\widehat{M}_W)$, where $\widehat{M}_W = \Sigma d_{w_i} \widehat{P}_{w_i}(\cdot \, ; W)$ is the* ***context-aware mixture representation*** *of W.*

Note, that our method can efficiently compute the surprisingness score for a new document without the need of retraining on the entire corpus, making it particularly suitable for many web applications.

EXPERIMENTS

To evaluate the proposed method, we used two datasets:

Interesting iPhone cases. Motivated by the important task of generating user engagement on an e-commerce website like eBay, we collected a number of interesting (positive) and uninteresting (negative) iPhone case product titles as follows. For generating positive examples, we used the data hosted by Pinterest. For generating the uninteresting iPhone case dataset (negative examples), we hired workers from *AMT* to label a collection of nearly 20,000 iPhone cases on *eBay*. We then pulled our final dataset from the annotated by selecting only those instances where the annotators all labeled it as uninteresting. The final dataset 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. We used the Mallet LDA implementation (see [11]) to learn a topic model with 400 topics.

Cross-disciplinary 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 (not surprising) or very different (surprising), based on the available meta-data, and labeling them accordingly. We generated 5,000 of those examples with positive and negative labels evenly represented. For this experiment, we trained a separate topic model with 300 topics based on the original NSF abstracts.

Unsupervised Setting

We evaluated our text surprisingness model in an unsupervised learning task on both datasets. The model was implemented as described in Sections 3 and 4 (labeled by *JSD-Sim-Con* in the plots). We compare out method against other known document diversity metrics: LDA topic entropy, LDA topic entropy using topic similarity (labeled by *Entropy-Sim*), *Rao diversity* (see [1] for details) and a metric based on determinantal point processes (DPP) [9], which measures the spread of word representations in a document as vectors in a Euclidean space.

As seen in Figures 3 (a) and (c), our approach outperforms the baselines, with an AUC around 0.73. Moreover, for the *eBay* dataset the other measures give very poor results. This can be explained as follows: since the text snippets are short, LDA may yield a poor topic inference for such short text and as a result all measures using topic inference directly on the test data would perform poorly. Our technique allows training LDA on a separate dataset, so it is not affected by text length. Interestingly, LDA topic entropy performs much better on the NSF dataset when applied to distributions transformed using topic similarity information, showing the efficacy of this technique proposed in Section 3. This is further confirmed in Figures 3 (b) and (d), which show gains obtained by separately applying topic similarity and context-awareness techniques to the word representations (see Sections 3 and 4), as well as using non-uniform importance weights d_w from Definition 2.

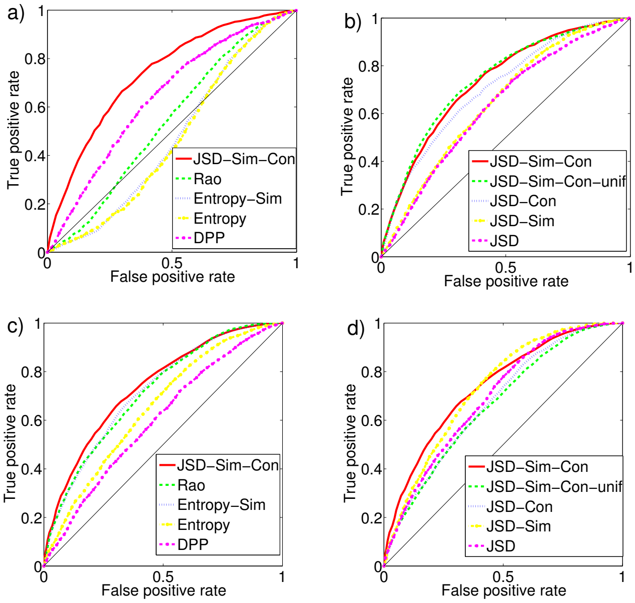


Figure 3. ROC curves presenting the results of experiments on the eBay dataset (a,b) and NSF proposal dataset (c,d). The comparison plots (a,c) show the results for our approach (JSD-Sim-Con) against other methods, while the plots (b,d) show different variations of our approach ("Sim" means applying topic similarity matrix S, "Con" means using context-aware word representations, rather than static ones, "unif" means uniform importance weights, instead of d_w).

Table 1. Classification results for the eBay dataset.

	Precision	Recall	Accuracy
JSD	**0.71** ± 0.01	0.60 ± 0.02	**0.883** ± 0.004
RAE	0.68 ± 0.01	**0.67** ± 0.03	0.881 ± 0.002
LSI	0.68 ± 0.01	0.63 ± 0.02	0.878 ± 0.003

Supervised Setting

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, see [5]), and a deep learning approach using the *recursive auto-encoders (RAE)* framework described in [17]. These results are averaged over five different cross-validation splits using 0.6 for training and 0.4 for testing. Our proposed approach shows marginally higher accuracy compared to the baselines, but it also achieves a significantly higher precision, which is especially important, given that the goal of this task is discovering interesting products for recommendation.

CONCLUSIONS

In this paper we propose a novel approach for discovering surprising documents, and show that it can be an effective way of discovering interesting products for recommendation. At the heart of this approach lies a model for constructing context-aware distributional word representations. Using this framework, we show that Jensen-Shannon Divergence can be a useful measure of text surprisingness. We provide experimental results in two different real world domains, for which this method outperforms the previously known metrics.

REFERENCES

1. Kevin Bache, David Newman, and Padhraic Smyth. 2013. Text-based Measures of Document Diversity. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, New York, NY, USA, 23–31.
2. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. `http://dl.acm.org/citation.cfm?id=944919.944937`
3. Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.
4. Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 107–116. `http://dl.acm.org/citation.cfm?id=1870568.1870582`
5. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landaue, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407.
6. B. Fuglede and F. Topsoe. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *IEEE International Symposium on Information Theory*. 31–31.
7. Debasis Ganguly, Johannes Leveling, and Gareth Jones. 2014. Automatic Prediction of Aesthetics and Interestingness of Text Passages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 905–916. `http://www.aclweb.org/anthology/C14-1086`
8. Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. Discovering diverse and salient threads in document collections. In *In EMNLP-12. Ralph Grishman, Catherine*.
9. Alex Kulesza and Ben Taskar. 2012. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA. `http://dl.acm.org/citation.cfm?id=2481023`
10. Igor Labutov and Hod Lipson. 2012. Humor as Circuits in Semantic Networks. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 150–155. `http://www.aclweb.org/anthology/P12-2030`
11. Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. (2002). http://mallet.cs.umass.edu.
12. Rada Mihalcea and Carlo Strapparava. 2005. 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)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 531–538.
13. C. Rao. 1982. Diversity and dissimilarity coefficients: a unified approach. In *Theoretical Population Biology*, Vol. 21(1). 24–43.
14. Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical population biology* 70, 3 (November 2006), 237–243.
15. Jürgen Schmidhuber. 2010. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). *IEEE T. Autonomous Mental Development* 2, 3 (2010), 230–247.
16. Ekaterina Shutova and Lin Sun. 2013. 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)*. Association for Computational Linguistics, 978–988. `http://aclweb.org/anthology/N13-1118`
17. Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 151–161. `http://dl.acm.org/citation.cfm?id=2145432.2145450`