Topic Modelling Games

Rocco Tripodi

Sapienza NLP Group

Department of Computer Science, Sapienza University of Rome

`tripodi@di.uniroma1.it`

Abstract

English. This paper presents a new topic modelling framework inspired by game theoretic principles. It is formulated as a *normal form* game in which words are represented as players and topics as strategies that the players select. The strategies of each player are modelled with a probability distribution guided by a utility function that the players try to maximize. This function induces players to select strategies similar to those selected by similar players and to choice strategies not shared with those selected by dissimilar players. The proposed framework is compared with state-of-the-art models demonstrating good performances on standard benchmarks.

Italiano. *Questo articolo presenta un approccio di modellazione dei topic ispirato alla teoria dei giochi. La modellazione dei topic è vista come un gioco in forma normale in cui le parole rappresentano i giocatori e i topic le strategie che i giocatori possono scegliere. Ogni giocatore sceglie le strategie da impiegare tramite una distribuzione di probabilità che viene influenzata da una funzione di utilità che i giocatori cercano di massimizzare. Questa funzione incentiva i giocatori a scegliere strategie simili a quelle impiegate da giocatori simili e disincentiva la scelta di strategie condivise con giocatori dissimili. Il confronto con modelli allo stato dell'arte dismostra buone prestazioni su diversi dataset di valutazione.*

1 Introduction

Topic modeling is a technique that discovers the underlying topics contained in a collection of documents (Blei, 2012; Griffiths and Steyvers, 2004). It can be used in different tasks of text classification, document retrieval, and sentiment analysis, providing together vector representations of words and documents. State-of-the-art systems are based on probabilistic (Blei et al., 2003; Mcauliffe and Blei, 2008; Chong et al., 2009) and neural networks models (Bengio et al., 2003; Hinton and Salakhutdinov, 2009; Larochelle and Lauly, 2012; Cao et al., 2015). A different perspective based on game theory is proposed in this article.

The use of game-theoretic principles in machine learning (Goodfellow et al., 2014), pattern recognition (Pavan and Pelillo, 2007) and natural language processing (Tripodi et al., 2016; Tripodi and Navigli, 2019) problems is developing a promising field of research with the development of original models. The main difference between computational models based on optimization techniques and game-theoretic models is that the former tries to maximize (minimize) a function (that in many cases is non-convex) and the latter tries to find the equilibrium state of a dynamical system. The equilibrium concept is useful because it represents a state in which all the constraints of a given system are satisfied and no object of the system has an incentive to deviate from it, because a different configuration will immediately lead to a worse situation in terms of payoff and fitness, at object and system level. Furthermore, it is guaranteed that the system converges to a mixed strategy Nash equilibrium (Nash, 1951). So far, game-theoretic models have been used in classification and clustering tasks (Pavan and Pelillo, 2007; Tripodi and Pelillo, 2017). In this work, it is proposed a game-theoretic model for inferring a low dimensional representation of words that can capture their latent semantic representation.

In this work, topic modeling is interpreted as a symmetric non-cooperative game (Weibull, 1997) in which, the words are the players and the topics

are the strategies that the players can select. Two players are matched to play the games together according to the co-occurrence patterns found in the corpus under study. The players use a probability distribution over their strategies to play the games and obtain a payoff for each strategy. This reward helps them to adjust their strategy selection in future games, considering what strategy has been effective in previous games. It allows concentrating more mass on the strategies that get high reward. The underlying idea to model the payoff function is to create two influence dynamics, the first one forces similar players (words that appear in similar contexts) to select similar strategies; the second one forces dissimilar players (words that do not share any context) to select different strategies. The games are played repeatedly until the system converges, that is, the difference among the strategy distributions of the players at time t and at time $t - 1$ is under a small threshold. The convergence of the system corresponds to an equilibrium, a situation in which there is an optimal association of words and topics.

2 Related Work

Hofmann (1999) proposed one of the earliest topic models, probabilistic Latent Semantic Indexing (pLSI). It represents each word in a document as a sample from a mixture model, where topics are represented as multinomial random variables and documents as a mixture of topics. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), the most widely used topic model, is a generalization of pLSI that introduces Dirichlet priors for both the word multinomial distributions over topics and topic multinomial distributions over documents. This line of research has been developed building on top of LDA different features to infer correlations among topics (Lafferty and Blei, 2006) or to model jointly words and labels in a supervised way (Mcauliffe and Blei, 2008).

Topic models based on neural network principles have been introduced with the neural network language model proposed in (Bengio et al., 2003). This paradigm is very popular in NLP and many topic models are based on it because with these techniques it is possible to obtain a low-dimensional representation of the data. In particular, auto-encoders (Ranzato and Szummer, 2008), Boltzmann machines (Hinton and Salakhutdinov, 2009) and autoregressive distributions (Larochelle

and Lauly, 2012) have been used to model documents with layer-wise neural network tools. Neural Topic Model (NTM; (Cao et al., 2015)) tries to overcome some limitations of classical topic models, such as the initialization problem and the generalization to n-grams. It exploits word embedding to represent n-grams and uses backpropagation to adjust the weights of the network between the embedding and the word-topic and document-topic layers. A general framework for topic modeling based also on neural networks is Sparse Contextual Hidden and Observed Language AutoencodeR (SCHOLAR; (Card et al., 2018)). It allows using covariates to influence the topic distributions and labels to include supervision. As Sparse Additive GEnerative models (SAGE; (Eisenstein et al., 2011))it can produce sparse topic representations but differently from it and Structural Topic Model (STM; (Roberts et al., 2014)) it can easily consider a larger set of metadata. A graphical topic model was proposed by Gerlach et al. (2018). In this framework, the task of finding topical structures is interpreted as the task of finding communities in complex networks. It is particularly interesting because it shows analogies with traditional topic models and overcomes some of their limitations such as the bound with a Bayesian prior and the need to specify the number of topics in advance.

3 Topic Modelling Games

Normal-form games consist of a finite set of players $N = (1, .., n)$, a finite set of pure strategies, $S_i = \{1, ..., m_i\}$ for each player $i \in N$ and a payoff (utility) function $u_i : S \rightarrow \mathbb{R}$, that associates a payoff to each combination of strategies $S = S_1 \times S_2 \times ... \times S_n$. The payoff function does not depend only on the strategy chosen by a single player but by the combination of strategies played at the same time by the players. Each player tries to maximize the value of u_i. Furthermore, in *non-cooperative games* the players choose their strategies independently, considering what other players can play and trying to find the best response to the strategy of the co-players. Nash equilibria (Nash, 1951) represent the key concept of game theory and can be defined as those strategy combinations in which each strategy is a best response to the strategy of the co-player and no player has the incentive to unilaterally deviate from them because there is no way to do better. In addition to play pure strategies, that correspond to select-

ing just one strategy from those available in S_i, a player i can also use *mixed strategies*, which are probability distributions over pure strategies. A mixed strategy over S_i is defined as a vector $\mathbf{x_i} = (x_1, \ldots, x_{m_i})$, such that $x_j \geq 0$ and $\sum x_j = 1$. In a *two-player game*, a strategy profile can be defined as a pair $(\mathbf{x}_i, \mathbf{x}_j)$. The expected payoff for this strategy profile is computed as:

$$u(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot A_{ij} \mathbf{x}_j$$

where A_{ij} is the $m_i \times m_j$ payoff matrix between player i and j.

Evolutionary game theory (Weibull, 1997) has introduced two important modifications: 1. the games are played repeatedly, and 2. the players update their mixed strategy over time until it is not possible to improve the payoff. The players, with these two modifications, can develop an inductive learning process, that allows them to learn their strategy distribution according to what other players are selecting. The payoff corresponding to the h-th pure strategy is computed as:

$$u(x_i^h) = x_i^h \cdot \sum_{j=1}^{n_i} (A_{ij}\mathbf{x}_j)^h \qquad (1)$$

The average payoff of player i is calculated as:

$$u(\mathbf{x}_i) = \sum_{h=1}^{m_i} u(x_i^h) \qquad (2)$$

To find the Nash equilibrium of the game, it is common to use the *replicator dynamics* equation (Weibull, 1997). It allows better than average strategies to grow at each iteration. It can be considered as an *inductive learning* process, in which the players learn from past experiences how to play their best strategy. It is important to notice that each player optimizes its individual strategy space, but this operation is done according to what other players simultaneously are doing so the local optimization is the result of a global process.

Data Preparation The players of the topic modelling games are the words $v = (1, \ldots, n)$ in the vocabulary V of the corpus under analysis and the strategies $S = (1, \ldots, m)$ are the topics to extract from the same corpus. The strategy space $\mathbf{x}_i$ of each player i is represented as a probability distribution that can be interpreted as the mixture of topics typically used in topic modeling. The interactions among the players are modeled using

the $n \times n$ adjacency matrix (W) of an undirected weighted graph. Each entry w_{ij} encodes the similarity between two words. The strategy space of the games can be represented as a $n \times m$ matrix X, where each row represents the probability distribution of a player over its m strategies (topics that have to be extracted from the corpus).

Payoff Function and System Dynamics The payoff function of the game is constructed exploiting the information stored in W. This matrix gives us the structural information of the corpus. It allows us to select the players with whom each player is playing the games, indicated with the presence of an edge between two nodes (players), and to quantify the level of influence that each player has on the other, indicated with the weight on each edge. The absence of an edge in this graph indicates that two words are distributional dissimilar. Using these three sources of information we model a payoff function that forces similar players to choose similar strategies (topics) and dissimilar players to choose different ones. The payoff of a player is calculated as,

$$u(x_i^h) = x_i^h (\sum_{j=1}^{n_i} (A_{ij}\mathbf{x}_j)^h - \sum_{g=1}^{neg_i} (\epsilon \mathbf{x}_g)^h) \qquad (3)$$

where the first summation is over all the n_i direct neighbors of player i that are the players with whom i share some similarity and the second summation is over the neg_i negative players of player i, that are players with whom player i does not share any similarity. With the first summation player i will negotiate with its neighbors a correlated strategy (topic), with the second he will deviate from the strategies chosen by negative players, this is done by subtracting the payoff that i would have gained if these negative players would have been his neighbors. The negative players are sampled from V according to frequency, in the same way, negative samples are selected in word embeddings models (Mikolov et al., 2013; Tripodi and Pira, 2017). The equation that gives us the probability of selecting a word as negative is:

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=0}^{n} f(w_j)^{3/4}}, \qquad (4)$$

where $f(w_i)$ is the frequency of word w_i. Since the similarity with negative players is 0 we introduced the parameter ϵ to weight their influence and set it to $(A > 0)$. The number of negative players,

neg_i, is set to n_i (number of neighbours of player i).

Once the players have played all the games with their neighbors and negative players, the average payoff of each player can be calculated with Equation (2). The payoff is higher when two words are highly correlated and have a similar mixed strategy. For this reason the *replicator dynamics* equation (Weibull, 1997) is used to compute the dynamics of the system. It pushes the players to be influenced by the mixed strategy of the co-players. This influence is proportional to the similarity between two players (A_{ij}). Once the influence dynamics do not affect the players the Nash equilibrium of the system is reached. The stopping criteria of the dynamics and are: 1. the maximum number of iterations (10^5); and 2. the minimum difference between two different iterations (10^{-3}) that is calculated as $\sum_{i=1}^{n} \mathbf{x_i}(t-1) - \mathbf{x_i}(t)$.

4 Experimental Results

In this section, we evaluate TMG and compare it with state-of-the-art systems.

4.1 Data and Setting

The datasets used to evaluate TMG are 20 Newsgroups[1] (20NG) and NIPS[2]. 20NG is a collection of about $20,000$ documents organized into 20 different classes. NIPS is composed of about $1,700$ NIPS conference papers published between 1987 and 1999 with no class information. Each text was tokenized and lowercased. The stop-words were removed and the vocabulary was constructed considering the 1000 and 2000 most frequent words in 20NG and NIPS, respectively. This choice is in line with previous work (Card et al., 2018). To keep the model as simple as possible, the *tf-idf* weighting was used to construct the feature vectors of the words and the cosine similarity was employed to create the adjacency matrix A. It is important to notice here that other sources of information can be easily included at this stage, derived from pre-trained word embeddings, syntactic structures or document metadata. Then A is sparsified taking only the r nearest neighbours of each node. r is calculated as $r = log(n)$ this operation reduces the computational cost of the algorithm and guarantees that the graph remains connected (Von Luxburg, 2007).

Dataset	TMG	SCHOLAR	NVDM	LDA
20NG	824	819	927	**791**
NIPS	1311	1370	1564	**1017**

Table 1: Comparison of the models as perplexity.

The strategy space of the players was initialized using a normal distribution to reduce the parameters of the framework[3]. The last two parameters of the systems concern the stopping criteria of the dynamics and are: 1. the maximum number of iterations (10^5); and 2. the minimum difference between two different iterations (10^{-3}) that is calculated as $\sum_{i=1}^{n} \mathbf{x_i}(t-1) - \mathbf{x_i}(t)$.

TMG has been compared with SCHOLAR[4], LDA[5] and NVDM[6]. We configured the NVDM network with two encoder layers (500-dimensional) and ReLu non-linearities. SCHOLAR has been configured using a more complex setting that consists in a single layer encoder and a 4-layer generator. LDA has been run with the following parameters: $\alpha = 50$, $iterations = 1000$ and $topic_{threshold} = 0$.

4.2 Evaluation

In this section, we compared the generalization performances of TMG and compared them with the models presented in the previous section. For the evaluation we used *perplexity* (PPL), even if it is has been shown to not correlate with human interpretation of topics (Chang et al., 2009). We computed perplexity on unobserved documents (**C**), as.

$$PPL(\mathbf{C}) = \exp(-\frac{1}{N}\frac{\sum_{n=1}^{N} logP(\mathbf{C}_n)}{\sum_{n=1}^{N} D_n}) \quad (5)$$

where N is the number of documents in the collection C. Low perplexity suggests less uncertainties about the documents. Held out documents represent the 15% of each dataset. Perplexity is computed for 10 topics for the NIPS dataset and 20 topics for the 20 Newsgroups dataset. These numbers correspond to the real number of classes of each dataset.

Table 1 shows the comparison of perplexity. As reported in previous work (Card et al., 2018), it is

---

[1] http://qwone.com/ jason/20Newsgroups/
[2] http://www.cs.nyu.edu/ roweis/data.html

[3] Experimentally it was also observed that using a Dirichlet distribution to initialize the strategy space with different α parameters did not affect much the performances of the model.
[4] https://github.com/dallascard/scholar
[5] http://mallet.cs.umass.edu
[6] https://github.com/ysmiao/nvdm

difficult to achieve a lower perplexity than LDA. The results in these experiments follow the same pattern, with LDA that has the lowest perplexity, TMG, and SCHOLAR that have similar results, and NVDM that performs slightly worse on both datasets.
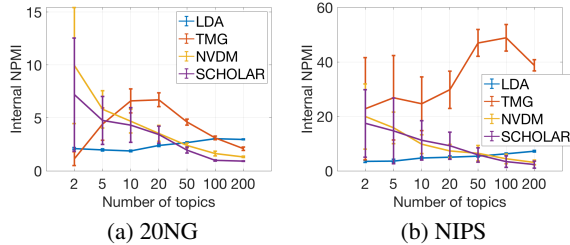

(a) 20NG

(b) NIPS

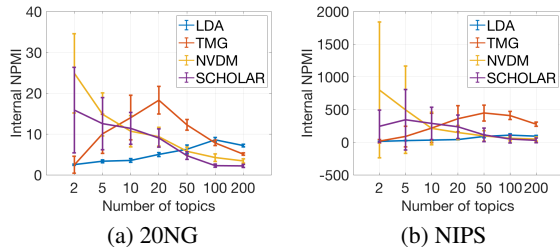Figure 1: Internal PMI mean and std values.

(a) 20NG

(b) NIPS

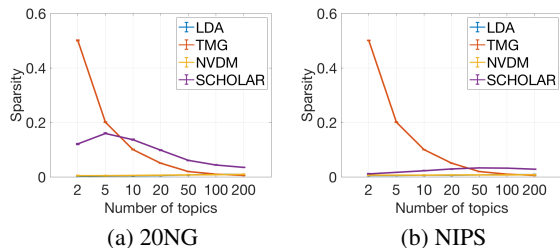Figure 2: External PMI mean and std values.

(a) 20NG

(b) NIPS

Figure 3: Sparsity mean and std values.

4.3 Topic Coherence and Interpretability

It has been shown that perplexity does not necessarily correlate well with topic coherence (Chang et al., 2009; Srivastava and Sutton, 2017). For this reason, we evaluated the performances of our system also on coherence (Chang et al., 2009; Das et al., 2015). The coherence is calculated by computing the relatedness between topic words using the pointwise mutual information (PMI). We used Wikipedia (2018.05.01 dump) as corpus to compute co-occurrence statistics using a sliding window of 5 words on the left and on the right of

each target word. For each topic, we selected the 10 words with the highest mass. Then we calculated the PMI among all the words pair and finally compute the coherence as the arithmetic mean of all these values. This metric has been shown to correlate well with human judgments (Lau et al., 2017). We used two different sources of information for the computation of the PMI: one is internal and corresponds to the dataset under analysis; the other one is external and is represented by the English Wikipedia corpus.

Internal PMI Figure 1 presents the PMI values of the different models computed on the two corpora. As it is possible to see from figure 1a, TMG has a low PMI compared to all other systems on the 20 Newsgroups dataset when there are few topics to extract (i.e.: 2 and 5). The situation changes drastically when the number of topics increases. In fact, it has the highest performances on this dataset when extracts 10, 20, 50, 100 topics. The performances of NDVM and SCHOLAR are similar and follow a decreasing pattern, with very high values at the beginning. On the contrary, the performances of LDA follow an opposite pattern this model seems to work better when the number of topics to extract is high. On NIPS (Figure 1b) the performances of the systems are similar to those on 20 Newsgroups. The only exception is that TMG has always the highest PMI and seems to behave better also when the number of topics to extract is high. This probably because the number of words in NIPS is higher and for this, it is reasonable to have also a higher number of topics. This is also confirmed from a qualitative analysis of the topics in Section 4.4, where it is demonstrated that with low values of k it is possible to extract general topics and increasing its value it is possible to extract more specific ones.

In general, we can find three different patterns in these experiments: 1. NDVM and SCHOLAR work well on extracting a low number of topics; 2. LDA works well when it has to extract a large number of topics; 3. TMG works well on extracting a number of topics that is close to the real number of classes in the datasets. Another aspect to take into account is the fact that even if TMG has the highest performances, its results have also a high standard deviation. This is due to the stochastic nature of negative sampling.

turks	schneider	drive	vms	god	intellect	bike	providing	fbi	gun	team	space	male	tim	amateur
soviet	allan	ide	disclaimer	jesus	banks	ride	encryption	compound	firearms	game	orbit	gay	israel	georgia
turkish	morality	scsi	vnews	christians	gordon	riding	clipper	batf	guns	play	shuttle	men	israeli	intelligence
armenian	keith	controller	vax	christ	surrender	dod	key	waco	criminals	season	launch	sexual	arab	ai
armenia	atheists	drives	necessarily	christianity	univ	bikes	escrow	children	crime	hockey	earth	percentage	jews	programs
passes	moral	mb	represents	bible	pittsburgh	motorcycle	crypto	koresh	weapons	league	mission	sex	arabs	michael
roads	political	disk	views	christian	significant	bmw	keys	criminal	criminal	nhl	flight	study	policy	radio
armenians	pasadena	isa	expressed	faith	hospital	honda	chip	gas	violent	players	nasa	apparent	war	adams
argic	objective	bus	news	church	level	road	secure	branch	weapon	cup	moon	showing	land	ignore
proceeded	animals	floppy	poster	belief	blood	advice	wiretap	started	armed	stanley	solar	women	north	occur
29.71	15.27	12.7	11.72	10.79	10.18	8.94	8.93	8.55	7.52	7.45	7.14	6.92	6.21	6.13

Table 2: Best topics (each topic is represented on the columns) extracted from 20 Newsgroup using TMG (setting $k = 20$) ordered using external PMI (bottom row).

ocular	dendrites	oscillatory	crowdsourcing	kaiming	retina	auditory	graph	disturbances	lifted
eye	dendritic	oscillations	crowds	shaoqing	photoreceptor	sound	edges	plant	propositional
fovea	soma	oscillators	workers	xiangyu	retinal	sounds	graphs	controllers	predicate
dominance	dendrite	oscillator	worker	jian	vertebrate	cochlear	optimisation	controller	grounding
saccades	axonal	oscillation	labelers	yangqing	schulten	ear	edge	disturbance	predicates
saccadic	axons	synchronization	crowd	karen	photoreceptors	hearing	vertices	plants	domingos
fixation	nmda	decoding	turk	sergey	ganglion	ears	optimise	activate	clauses
foveal	pyramidal	locking	wisdom	trevor	kohonen	acoust	optimising	activated	compilation
eyes	somatic	synchronize	expertise	sergio	bipolar	tone	optimised	activating	formulas
saccade	axon	synchronized	dawid	jitendra	visualizing	cochlea	vertex	activates	logical
304.85	283.66	276.39	230.5	218.51	196.86	176.75	146.3	146.25	145.84

Table 3: Topics extracted from NIPS using TMG (setting $k = 10$) ordered using external PMI (bottom row).

Sparsity We compared the sparsity of the word-topics matrices, X, in Figure 3a and 3b, computed as $s = \frac{|X > 10^{-3}|}{|X|}$. From both figures, we can see that TMG can produce highly sparse representations especially when the number of topics to extract is low. This is a nice feature since it provides more interpretable results. Only SCHOLAR produces more sparse representations when the number of topics to extract is high. Experimentally we also noticed that we can control the sparsity of X, in TMG, increasing the number of iterations of the game dynamics.

4.4 Qualitative Evaluation

Examples of topics extracted from 20NG and NIPS are presented in Table 2 and 3, respectively[7]. The first difference that emerges from these results are the external PMI values. This is due to the fact that the texts in NIPS have a very specific language and for this reason the PMI values are very high. We can also see that TMG groups highly coherent set of words in each topic. We can easily identify in Table 2 the topics in which the dataset is organized and especially: talk.politics.midleast, alt.atheism, comp.graphics, soc.religion.christian, talk.politics.misc, rec.motorcycles, sci.crypt, talk.politics.guns, rec.sport.hockey, sci.space, talk.politics.misc.

We can also easily identify from Table 3 highly coherent topics, related to optic, signal analysis, optimization, crowdsourcing, audio, graph theory and logics. We noticed from these topics that they are general and that it is possible to discover more specific topics increasing the number of topics to extract. For example, we discovered topics related to topic modelling and generative adversarial networks.

5 Conclusion and Future Work

In this paper, it is presented a new topic modeling framework based on game-theoretic principles. The results of its evaluation show that the model performs well compared to state-of-the-art systems and that it can extract topically and semantically related groups of words. In this work, the model was left as simple as possible to assess if a game-theoretic framework itself is suited for topic modeling. In future work, it will be interesting to introduce the topic-document distribution and to test it on classification tasks and covariates to extract topics using different dimensions, such as time, authorship, or opinion. The framework is open and flexible and in future work, it will be tested with different initializations of the strategy space, graph structures, and payoff functions. It will be particularly interesting to test it using word embedding and syntactic information.

---

[7]for space limitation we presented only 15 topics for 20NG

References

[Bengio et al.2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Blei2012] David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April.

[Cao et al.2015] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *AAAI*, pages 2210–2216.

[Card et al.2018] Dallas Card, Chenhao Tan, and Noah A Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the ACL*, volume 1, pages 2031–2040.

[Chang et al.2009] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296.

[Chong et al.2009] Wang Chong, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *CVPR, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE.

[Das et al.2015] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the ACL*, volume 1, pages 795–804.

[Eisenstein et al.2011] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text.

[Gerlach et al.2018] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. 2018. A network approach to topic models. *Science Advances*, 4(7).

[Goodfellow et al.2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.

[Griffiths and Steyvers2004] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[Hinton and Salakhutdinov2009] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *NIPS*, pages 1607–1614.

[Hofmann1999] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pages 50–57. ACM.

[Lafferty and Blei2006] John D Lafferty and David M Blei. 2006. Correlated topic models. In *NIPS*, pages 147–154.

[Larochelle and Lauly2012] Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. In *NIPS*, pages 2708–2716.

[Lau et al.2017] Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the ACL*, volume 1, pages 355–365.

[Mcauliffe and Blei2008] Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *NIPS*, pages 121–128.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[Nash1951] John Nash. 1951. Non-cooperative games. *Annals of mathematics*, pages 286–295.

[Pavan and Pelillo2007] Massimiliano Pavan and Marcello Pelillo. 2007. Dominant sets and pairwise clustering. *IEEE transactions on pattern analysis and machine intelligence*, 29(1).

[Ranzato and Szummer2008] Marc'Aurelio Ranzato and Martin Szummer. 2008. Semi-supervised learning of compact document representations with deep networks. In *Proceedings of the 25th international conference on Machine learning*, pages 792–799. ACM.

[Roberts et al.2014] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

[Srivastava and Sutton2017] Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations (ICLR)*.

[Tripodi and Navigli2019] Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 88–99, Hong Kong, China, November. Association for Computational Linguistics.

[Tripodi and Pelillo2017] Rocco Tripodi and Marcello Pelillo. 2017. A game-theoretic approach to word sense disambiguation. *Computational Linguistics*, 43(1):31–70.

[Tripodi and Pira2017] Rocco Tripodi and Stefano Li Pira. 2017. Analysis of italian word embeddings. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*.

[Tripodi et al.2016] Rocco Tripodi, Sebastiano Vascon, and Marcello Pelillo. 2016. Context aware nonnegative matrix factorization clustering. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 1719–1724.

[Von Luxburg2007] Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[Weibull1997] J. W. Weibull. 1997. *Evolutionary game theory*. MIT press.