Minority report detection in refugee-authored community-driven journalism using RBMs

Bogdana Rakova
Think Tank Team
Samsung Research America
665 Clyde Ave, Mountain View, CA
b.rakova@samsung.com

Nick DePalma
Artificial Intelligence Center
Samsung Research America
665 Clyde Ave, Mountain View, CA
n.depalma@samsung.com

Abstract

Our work seeks to gather and distribute sensitive information from refugee settlements to stakeholders to help shape policy and help guide action networks. In this paper, we propose the following 1) a method of data collection through stakeholder organizations experienced in working with displaced and refugee communities, 2) a method of topic modeling based on Deep Boltzmann Machines that identifies topics and issues of interest within the population, to help enable mapping of human rights violations, and 3) a secondary analysis component that will use the probability of fit to isolate minority reports within these stories using anomaly detection techniques.

1 Introduction

The United Nations Sustainable Development Goals, or SDGs(1), are a set of seventeen objectives that must be addressed in order to bring the world to a more equitable, prosperous, and sustainable path. Our work is focused on how new types of AI tools could help promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels(2). In our analysis and proposal we focus on those fleeing violence, turmoil, or seeking opportunity in a foreign country where they are often denied their basic human rights and liberties(3).

Expressive journaling techniques have frequently been employed to moderate strong emotional experience and to aid in coping(4). Large N trials have demonstrated that journaling can help high anxiety patients lower blood pressure, reduce recurring anxious thoughts, and cope with traumatic experience. While we do not expect these effects to compensate for the experiences refugees go through, we do believe that allowing those refugees to express themselves in a safe environment is mutually beneficial to all stakeholders. Our hypothesis is that, *many journal entries will share subject matter, major events, and potential solutions* within the settlement. Using topic modelling, we can track these frequently told stories. But while major events like publicly viewed human rights violations may be frequently clustered to a topic, lesser and more intimate violations such as sexual abuse may not be modeled as well in standard topic models.

We present an unsupervised study of journal articles using Deep Boltzmann Machines(DBMs)(5). We analyze the potential implications of the development of community-driven journalism tools in the broader ecosystems of organizations and individuals working to provide sustainable solutions to the problems facing displaced communities. We discuss foreseeable challenges, threat models and next steps.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

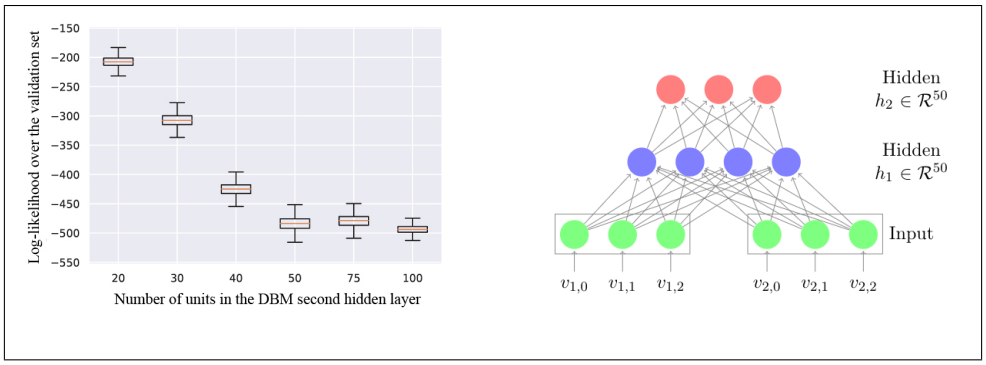

Figure 1: DBM hidden layer architecture analysis(left) and our Replicated Softmax RBM per document model(right).

2 Related work

Making long lasting impact requires coordination with organizations with impact and reach to the populations we are most interested in. Microsoft's Project Fortis(6) is one such project that dynamically aggregates data and enables data scientists to perform statistical analysis collected from popular social media and humanitarian databases like those published from OCHA(7). We differ from this system in that we are interested in proposing models that identify minority reports using meta analysis of documents originating from refugee settlements. Rather than build tools to aggregate data, we hope to use AI techniques to reveal patterns and highlight minority reports that need attention and resources.

Deep Learning has recently been applied in the context of coordinating humanitarian relief efforts in conflict situations through the analysis of remote sensing imagery from satellites or drones(8). Our work develops a complimentary approach that highlights the importance of community authored stories. Topic modeling is a key component in the fields of information retrieval(9) and understanding(10).

We build on the analysis of the main challenges posed by Hu et al.(11) and focus on the way non-experts can use topic modeling tools to aid their needs. We differ from their system as our topic modeling approach uses a different methodology that potentially would allow the system to reason about anomalies as well as analyze multimodal data(12).

3 Story collection

The goal of our work is to analyze if AI-enabled topic modeling and anomaly detection can aid the work of humanitarian action and advocacy groups, and investigative journalists(13) by helping them organize different pieces of evidence. We begin by acknowledging that a transparent data collection process is crucial for the real-world success of any potential proposal. It must preserve peoples' privacy and must be informed by all involved stakeholders. For the purpose of our experiments, we use a database consisting of 6,258 news articles published by major media outlets between 01/2016 - 09/2017 and mentioning the refugee crisis(14). In future work we aim to collaborate directly with practitioners(15; 16; 17; 18) to help them extract and map insights from case work related to human rights violations in refugee settlements.

4 Methodology and preliminary results

We start by training a Deep Boltzmann Machine model (DBM) - a two-hidden layer Replicated Softmax RBM(19) with weights sharing. The visible units in our architecture correspond to lemmatized word count vectors where each vector represents a document in the training corpus(see Figure 1, right). As with a standard RBMs the learning proceeds via Contrastive Divergence(20).

We hypothesize that the DBM model learns the peculiarities that are more prevalent in the training data, producing better reconstructions of them. We hypothesize that minority report entries will

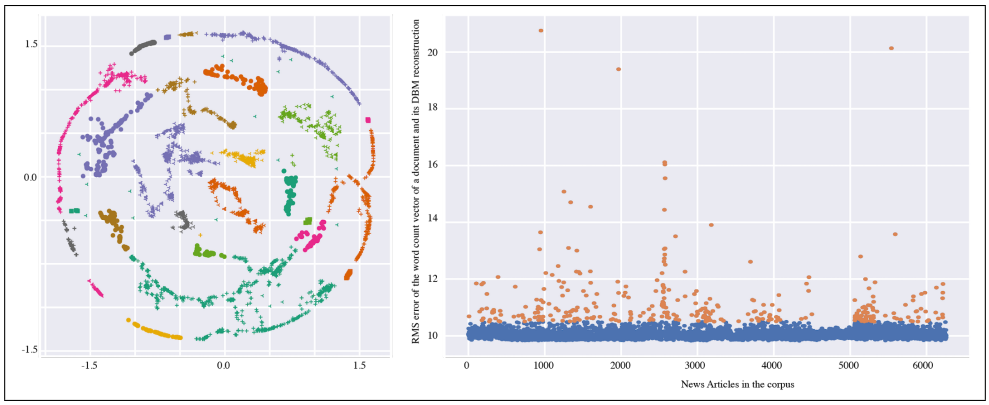

Figure 2: The results from applying DBSCAN clustering of the DBM representations of all articles in the corpus(left) and minority report selection using our metric. Selected minority reports are highlighted in orange(right).

occur rarely in the provided training data, preventing the DBM from learning and reconstructing it. Therefore we develop a metric for discovering these minority reports by measuring the distance between a document's input vector and its reconstruction. We aim to highlight those documents so that social scientists interacting with the algorithmic system could leverage these information retrieval insights. Following training of the traditional DBM parameters, $W_{1,i}, W_2, b_{1,i}, b_2$ corresponding to the weights and biases found for document i, we compute the reconstruction error. We first define the forward-backward pass mapping as follows:

$$\tilde{v}_i = \sigma(W_2 \cdot \sigma(W_1 \cdot v_i + b_1) + b_2) \tag{1}$$

$$\hat{v}_i = \sigma(W_1 \cdot \sigma(\tilde{v}_i \cdot W_2 + b_2') + b_1') \tag{2}$$

$$\epsilon(v_i, \hat{v}_i) = |\hat{v}_i - v_i| \tag{3}$$

Where v_i is the word count vector from a predefined dictionary for a specific document i, $\tilde{v}_i$ is the latent hidden embedding of that document from the DBM and $\hat{v}_i$ is the reconstruction of v_i given $\tilde{v}_i$. Finally σ is the non-linear activation function, $\sigma(x) = \frac{1+\tanh(\frac{x}{2})}{2}$. The dictionary consisted of the top 1000 most common terms in the training corpus. The weights and biases represent the symmetric interaction terms between visible-to-hidden and hidden-to-hidden variables learned while pre-training the Softmax RBMs and fine-tuning the DBM by minimizing the negative log-likelihood over a holdout validation set(Figure 1, left).

We have performed an initial investigation into our trained model by first generating the latent embeddings $\tilde{v}$ of all the documents in the corpus and clustering the results using DBSCAN, an agglomerative clustering technique.

We visualized the results in Figure 2, left using a t-SNE visualization(21). By analyzing word frequencies in each cluster we find that for more than half of all clusters the US President Trump is a major actor, however individual subclusters were formed related to other political figures such as the US Secretary of State John Kerry and Germany's Chancellor Angela Merkel. Another subcluster emerges related to terms such as women, public health and policy. Finally, we use our minority report metric to extract the distance between a document and its reconstruction and plot the result in Figure 2, right. As discussed previously, we believe that the outliers indicate that some of these documents were not modeled well and that their reconstruction error would be significantly larger than the other documents in the corpus.

Initial probing of these minority reports show that they cover the meta-topics of: 1) articles not related to the refugee crisis at all, 2) articles expressing specific sentiment which we don't see often in media, 3) articles which were very concrete and graphic about violence, 4) articles about the Cold War, and 5) an assortment of other topics.

5 Challenges and future work

Violent conflicts and natural disasters are causing large numbers of civilian casualties(2). Communities bring the energy and expertise to reinvent themselves from within. They know what they need more than anybody that we could possibly bring in from the outside. Algorithmic tools could aid multi-lingual cross-cultural understanding however we need a broader conversation where vulnerable communities are empowered to participate.

We understand that sensitive data in refugee settlements can be used for ill purposes and would like to highlight the importance of future work in the fields of privacy, trust and secure information sharing systems. Loss, theft, abuse, misuse, and unintended actions with datasets threaten the lives of these individuals and may lead to irreversible consequences(22).

It is important to consider that any algorithmic tool we create will be used in unintended ways and therefore we need to create threat models that identify vulnerabilities and define countermeasures to prevent, or mitigate the effects of threats. Through conversations with stakeholders, we have also found that the expected number of documents are far fewer than the dataset that was used for our experiment. Future work should explore how to perform the same type of analysis with fewer examples. Furthermore, could technology help us detect artificially crafted and fake data designed to make people believe and act in certain ways? We need to ask these questions and rigorously study the implications of the available technological tools in the broader social context where they are being applied.

6 Acknowledgements

This work was greatly influenced by Brent Dixon, co-organizer and chair of Greece Communitere - the Greece chapter of an international NGO creating dynamic, collaborative hubs in displaced and post-disaster communities. We would also like to thank the generosity of Samsung Research for encouraging our investigation into these issues.

References

[1] United Nations, "UN Sustainable Development Goals." https://sustainabledevelopment.un.org.

[2] United Nations, "UN Sustainable Development Goal 16." https://sustainabledevelopment.un.org/sdg16.

[3] V. Bollettino, S. Campo, S. Campo, K. Crawford, S. McDonald, Martin, and M. Whittaker, "The signal code: A human rights approach to information during crisis," *Harvard Humanitarian Initiative*, 2017.

[4] A. N. Niles, K. E. B. Haltom, C. M. Mulvenna, M. D. Lieberman, and A. L. Stanton, "Randomized controlled trial of expressive writing for psychological and physical health: the moderating role of emotional expressivity," *Anxiety, Stress & Coping*, vol. 27, no. 1, pp. 1–17, 2014.

[5] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (D. van Dyk and M. Welling, eds.), vol. 5 of *Proceedings of Machine Learning Research*, (Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA), pp. 448–455, PMLR, 16–18 Apr 2009.

[6] E. Schlegel, "Project fortis: Accelerating un humanitarian aid planning with graphql," Jun 2017.

[7] United Nations, "The Humanitarian Data Exchange." https://data.humdata.org/.

[8] J. A. Quinn, M. M. Nyhan, C. Navarro, D. Coluccia, L. Bromley, and M. Luengo-Oroz, "Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, 2018.

[9] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, (New York, NY, USA), pp. 178–185, ACM, 2006.

[10] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, (Stroudsburg, PA, USA), pp. 363–371, Association for Computational Linguistics, 2008.

[11] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine Learning*, vol. 95, pp. 423–469, Jun 2014.

[12] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via dirichlet forest priors," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, (New York, NY, USA), pp. 25–32, ACM, 2009.

[13] R. J. Tofel, "Non profit journalism issues around impact." https://s3.amazonaws.com/propublica/assets/about/LFA_ProPublica-white-paper_2.1.pdf, 2018.

[14] A. Thompson, "All the news challenge by kaggle." https://www.kaggle.com/snapcrack/all-the-news, 2017.

[15] "Velos - provide support and protection to vulnerable populations." https://velosyouth.org/, 2017.

[16] "Advocates abroad - a legal advocacy group." https://advocatesabroad.org/, 2017.

[17] "Campfire Innovation - humanitarian grassroot organization." https://campfireinnovation.org, 2016.

[18] "The refugee journalism project by london college of communication." http://migrantjournalism.org/, 2016.

[19] G. E. Hinton and R. R. Salakhutdinov, "Replicated softmax: an undirected topic model," in *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 1607–1614, Curran Associates, Inc., 2009.

[20] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] M. Latonero, D. Poole, and J. Berens, "Refugee connectivity: A survey of mobile phones, mental health, and privacy at a syrian refugee camp in greece," *Harvard Humanitarian Initiative*, April 2018.