Appendix C TIME TRAINING

In this section, we present the time spent to train each embeddings and the two Deep Learning techniques for the 9 datasets.

Let's start by looking at the amount of time in seconds/minutes to train each embedding. Thus, in this experiment we will measure the time spent for each technique to build a embedding and train the logistic regression estimator. Results can be analyzed in the table 7.

Technique	Dataset	Time(s)
Word2Vec	SIM	4.01
Glove	SIM	4.01
Bayesian Sp	SIM	14.01
FastText	SIM	47.0
DiVe Dual	SIM	3.1
Word2Vec	AR	4.01
Glove	AR	4.01
Bayesian Sp	AR	9.01
FastText	AR	41.0
DiVe Dual	AR	4.01
Word2Vec	YR	4.01
Glove	YR	4.01
Bayesian Sp	YR	9.01
FastText	YR	47.0
DiVe Dual	YR	4.01
Word2Vec	CR	17.01
Glove	CR	19.01
Bayesian Sp	CR	33.0
FastText	CR	47.0
DiVe Dual	CR	15.0
Word2Vec	QTS	36.01
Glove	QTS	37.01
Bayesian Sp	QTS	45.0
FastText	QTS	69.0
DiVe Dual	QTS	30.0
heightTechnique	Dataset	Time(M)
Word2Vec	HSTW	2:14
Glove	HSTW	3:05
Bayesian Sp	HSTW	3:24
FastText	HSTW	2:50
DiVe Dual	HSTW	5:48
Word2Vec	PL	1:05
Glove	PL	1:45
Bayesian Sp	PL	2:27
FastText	PL	1:48
DiVe Dual	PL	1:41
Word2Vec	SUBJ	1:16
Glove	SUBJ	1:34
Bayesian Sp	SUBJ	2:14
FastText	SUBJ	1:39
DiVe Dual	SUBJ	1:22
Word2Vec	IM	15:48
Glove	IM	8:52
FastText	IM	17:48
DiVe Dual	IM	7:48

Table 7. Time(minutes or seconds) spent to train word embedding

From the table 7 we can see that all embeddings have very close training time, and that this time is relative to the dataset size. In addition, we note that the FastText and Dive techniques

generally have more training time than the others. Now Let's look at time spent to train deep learning techniques.

Technique	Dataset	Time(m)
ELMo	SIM	02:15
BERT	SIM	01:47
ELMo	AR	02:55
BERT	AR	01:17
ELMo	YR	2:14
BERT	YR	01:45
ELMo	QTS	08:10
BERT	QTS	08:13
ELMo	CR	05:47
BERT	CR	04:36
ELMo	HSTW	20:51
BERT	HSTW	12:52
ELMo	PL	14:33
BERT	PL	10:37
ELMo	SUBJ	12:41
BERT	SUBJ	09:44
ELMo	IM	40:57
BERT	IM	44:11

Table 8. Time(minutes) spend to train deep learning techniques

Thus, from the analysis of the tables, we can notice that the deep learning present much longer training time than word embeddings. In many cases, this time is about 10 times longer (YES, AR, CR, YR) and about 4 times longer (IM, SUBJ, HSTW).

Therefore, we can conclude that while BERT and ELMo enable greater accuracy, these techniques require a longer training time, often about 10 times longer, besides that, as we saw in our paper, this longer time spent does not necessarily reflect higher Accuracy.

REFERENCES

[1] 2009. SVD Transformation. In *Encyclopedia of Database Systems*. 2896. https://doi.org/10.1007/978-0-387-39940-9_3726

[2] Jacob Andreas and Dan Klein. 2015. When and why are log-linear models self-normalizing?. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. 244–249. http://aclweb.org/anthology/N/N15/N15-1027.pdf

[3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. *CoRR* abs/1706.00188 (2017). arXiv:1706.00188 http://arxiv.org/abs/1706.00188

[4] Mikhail Belkin and Partha Niyogi. 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* 15, 6 (June 2003), 1373–1396. https://doi.org/10.1162/089976603321780317

[5] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *JOURNAL OF MACHINE LEARNING RESEARCH* 3, 1137–1155.

[6] Y. Bengio and J. S. Senecal. 2008. Adaptive Importance Sampling to Accelerate Training of a Neural Probabilistic Language Model. *Trans. Neur. Netw.* 19, 4 (April 2008), 713–722. https://doi.org/10.1109/TNN.2007.912312

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* abs/1607.04606 (2016). arXiv:1607.04606 http://arxiv.org/abs/1607.04606

[8] Arthur Brazinskas, Serhii Havrylov, and Ivan Titov. 2018. Embedding Words as Distributions with a Bayesian Skip-gram Model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 1775–1789.

[9] Samuel Brody and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*. 103–111. http://www.aclweb.org/anthology/E09-1013

[10] Ethan Caballero. 2015. Skip-Thought Memory Networks. *CoRR* abs/1511.06420 (2015). arXiv:1511.06420 http://arxiv.org/abs/1511.06420

[11] Shaosheng Cao and Wei Lu. 2017. Improving Word Embeddings with Convolutional Feature Learning and Subword Information. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 3144–3151. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14724

[12] Shuo Chen, J. Moore, D. Turnbull, and T. Joachims. 2012. Playlist Prediction via Metric Embedding. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 714–722.

[13] Shuo Chen, Jiexun Xu, and T. Joachims. 2013. Multi-space Probabilistic Sequence Modeling. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 865–873.

[14] Wenlin Chen, David Grangier, and Michael Auli. 2015. Strategies for Training Large Vocabulary Neural Language Models. *CoRR* abs/1512.04906 (2015). arXiv:1512.04906 http://arxiv.org/abs/1512.04906

[15] Jian Cheng and Marek J. Druzdzel. 2011. AIS-BN: An Adaptive Importance Sampling Algorithm for Evidential Reasoning in Large Bayesian Networks. *CoRR* abs/1106.0253 (2011). arXiv:1106.0253 http://arxiv.org/abs/1106.0253

[16] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *EMNLP*.

[17] Trevor F. Cox and M.A.A. Cox. 2000. *Multidimensional Scaling*. Chapman and Hall/CRC.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[19] Chris Dyer. 2014. Notes on Noise Contrastive Estimation and Negative Sampling. *CoRR* abs/1410.8251 (2014). arXiv:1410.8251 http://arxiv.org/abs/1410.8251

[20] Manaal Faruqui and Chris Dyer. 2014. Community Evaluation and Exchange of Word Vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. 19–24. http://aclweb.org/anthology/P/P14/P14-5004.pdf

[21] Flavio Figueiredo, Bruno Ribeiro, Jussara M. Almeida, and Christos Faloutsos. 2015. TribeFlow: Mining & Predicting User Trajectories. *CoRR* abs/1511.01032 (2015). arXiv:1511.01032 http://arxiv.org/abs/1511.01032

[22] Peng Fu, Zheng Lin, Fengcheng Yuan, Weiping Wang, and Dan Meng. 2018. Learning Sentiment-Specific Word Embedding via Global Sentiment Representation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. 4808–4815. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16334

[23] Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. 2007. Euclidean Embedding of Co-occurrence Data. *J. Mach. Learn. Res.* 8 (Dec. 2007), 2265–2295. http://dl.acm.org/citation.cfm?id=1314498.1314572

[24] Yoav Goldberg and Graeme Hirst. 2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

[25] Michael U. Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *J. Mach. Learn. Res.* 13, 1 (Feb. 2012), 307–361. http://dl.acm.org/citation.cfm?id=2503308.2188396

[26] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. *CoRR* abs/1602.03483 (2016). arXiv:1602.03483 http://arxiv.org/abs/1602.03483

[27] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[28] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On Using Very Large Target Vocabulary for Neural Machine Translation. *CoRR* abs/1412.2007 (2014). arXiv:1412.2007 http://arxiv.org/abs/1412.2007

[29] Peng Jin, Yue Zhang, Xingyuan Chen, and Yunqing Xia. 2016. Bag-of-embeddings for Text Classification. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2824–2830. http://dl.acm.org/citation.cfm?id=3060832.3061016

[30] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *CoRR* abs/1612.03651 (2016). arXiv:1612.03651 http://arxiv.org/abs/1612.03651

[31] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *CoRR* abs/1607.01759 (2016). arXiv:1607.01759 http://arxiv.org/abs/1607.01759

[32] Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (1st ed.). Prentice Hall PTR, Upper Saddle River, NJ, USA.

[33] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2015. Character-Aware Neural Language Models. *CoRR* abs/1508.06615 (2015). arXiv:1508.06615 http://arxiv.org/abs/1508.06615

[34] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From Group to Individual Labels Using Deep Features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 597–606. https://doi.org/10.1145/2783258.2783380

[35] Rémi Lebret and Ronan Collobert. 2014. Rehabilitation of Count-based Models for Word Vector Representations. *CoRR* abs/1412.4930 (2014). arXiv:1412.4930 http://arxiv.org/abs/1412.4930

[36] Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based Neural Machine Translation. *CoRR* abs/1511.04586 (2015). arXiv:1511.04586 http://arxiv.org/abs/1511.04586

[37] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 142–150. http://dl.acm.org/citation.cfm?id=2002472.2002491

[38] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

[39] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. *CoRR* abs/1708.00107 (2017). arXiv:1708.00107 http://arxiv.org/abs/1708.00107

[40] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *CoRR* abs/1310.4546 (2013). arXiv:1310.4546 http://arxiv.org/abs/1310.4546

[42] Ricardo Pio Monti, Romy Lorenz, Robert Leech, Christoforos Anagnostopoulos, and Giovanni Montana. 2016. Text-mining the neurosynth corpus using deep boltzmann machines. In *International Workshop on Pattern Recognition in Neuroimaging, PRNI 2016, Trento, Italy, June 22-24, 2016*. 1–4. https://doi.org/10.1109/PRNI.2016.7552329

[43] J. Moore, Shuo Chen, T. Joachims, and D. Turnbull. 2012. Learning to Embed Songs and Tags for Playlist Prediction. In *Conference of the International Society for Music Information Retrieval (ISMIR)*. 349–354.

[44] J. L. Moore, T. Joachims, and D. Turnbull. 2014. Taste Space Versus the World: an Embedding Analysis of Listening Habits and Geography. In *Conference of the International Society for Music Information Retrieval Conference (ISMIR)*. 439–444.

[45] Frederic Morin and Yoshua Bengio. 2005. Hierarchical Probabilistic Neural Network Language Model. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, Robert G. Cowell and Zoubin Ghahramani (Eds.). Society for Artificial Intelligence and Statistics, 246–252. http://www.iro.umontreal.ca/~lisa/pointeurs/hierarchical-nnlm-aistats05.pdf

[46] Aida Nematzadeh, Stephan C. Meylan, and Thomas L. Griffiths. 2017. Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. In *CogSci*.

[47] Nouha Othman, Rim Faiz, and Kamel Smaili. 2017. A Word Embedding based Method for Question Retrieval in Community Question Answering. In *ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing*. ISGA, Casablanca, Morocco. https://hal.inria.fr/hal-01660005

[48] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543. https://nlp.stanford.edu/pubs/glove.pdf

[49] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[50] Lawrence R. Rabiner. 1990. Readings in Speech Recognition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, 267–296. http://dl.acm.org/citation.cfm?id=108235.108253

[51] Alec Radford. 2018. Improving Language Understanding by Generative Pre-Training.

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. (2018). https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units. *CoRR* abs/1508.07909 (2015). arXiv:1508.07909 http://arxiv.org/abs/1508.07909

[54] Yikang Shen, Wenge Rong, Nan Jiang, Baolin Peng, Jie Tang, and Zhang Xiong. 2015. Word Embedding based Correlation Model for Question/Answer Matching. *CoRR* abs/1511.04646 (2015). arXiv:1511.04646 http://arxiv.org/abs/1511.04646

[55] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[56] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*. 1067–1077. https://doi.org/10.1145/2736277.2741093

[57] J. B. Tenenbaum, V. Silva, and J. C. Langford. 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 5500 (2000), 2319–2323.