Appendix B DATASETS RELATED WORK

In this section, we will expose some studies in the literature that worked with the datasets presented by this work. Thus, in this section we will expose some techniques used by these works and their results, and how these results can be compared with that presented by this work.

B.1 Hate Speech Twitter Annotations (HSTW)

For this dataset, we selected two works in the literature that studied this dataset. In the work of [62] is used logistic regression with features based on n-grams, gender and location, trained in 10 folds and precision measured by the F1 metric. The results of this work can be found in Table 4. In the [3] work, two Deep Learning techniques, DNN, CNN and LSTM, with features based on pre-trained word embedding, TF-IDF and bag of words are studied. In this work, 10 folds were used to train the techniques and metric F1 score was used, table 4.

Technique	F1 score	Work
n-grams	0.7389	[62]
n-grams+gender	0.7393	[62]
n-grams+location	0.7362	[62]
DNN's+LSTM+Glove	0.808	[3]
DNN's+FastText+Glove	0.82	[3]
DNN's+FastText+RandomEmbedding	0.825	[3]
DNN's+LSTM+GloVe+GBDT	0.848	[3]
DNN's+LSTM+Random Embedding+GBDT	0.93	[3]

Table 4. Datasets used in classification tasks.

From the results from table 4, we can see that several results were surpassed by DiVe, which presents, in general, a simpler approach to training. Moreover, even the combination of 2 or 3 techniques ([3]) did not allow in some scenarios better accuracy than DiVe. As is also worth noting, performing 10-fold training enables a larger training set and smaller test set than 5-fold training, allowing for this change to achieve better accuracy.

B.2 Customer Review(CR), Question Type Classification (QTS) ,Polarity Opinion(PO) e IMDB reviews (IM)

For these datasets, we highlight the works of [9], [16] and [70] that used Deep Learning techniques, pre-trained word embedding, and Transfer Learning to classify these datasets. In these works, there was a separation of the examples of these datasets in training, test validation. No accuracy metric was used, so simple precision was used. The results can be seen in table 5.

From table 5, we can see that the accuracy found by these works is comparable to the accuracy found by DiVe, although our technique is using a metric that penalizes the imbalance between classes. As an example, in the QTS Dataset which has some unbalance between its 6 classes DiVe achieves values greater than 0.95 in simple precision.

B.3 Yelp reviews(YR), Small IMDB reviews(SIM) e Amazon reviews (AR)

For these 3 datasets, [34] 's work used a CNN - based architecture using a new objective function that considers the similarity between the instances. Results can be found in table 6.

Technique	Dataset	Acurácia	Work
AdaSent	QTC	0.9219	[70]
LSTM-RNN (without transfer)	QTS	0.9019	[9]
LSTM-RNN (with transfer)	QTS	0.93	[9]
LSTM-RNN (with transfer)	QTS	0.93	[9]
BiLSTM-Max(with transfer)	QTS	0.90	[16]
AdaSent	CR	0.8384	[70]
BiLSTM-Max(with transfer)	CR	0.69	[16]
AdaSent	PO	0.7984	[70]
LSTM-RNN (without transfer)	PO	0.750	[9]
LSTM-RNN (with transfer)	PO	0.80	[9]
BiLSTM-Max(with transfer)	PO	0.69	[16]
LSTM-RNN (without transfer)	IM	0.87	[9]

Table 5. Datasets used in classification tasks.

Technique	Dataset	Acurácia	Work
GICF w/Embeddings	SIM	0.86	[34]
GICF w/Embeddings	YR	0.86	[34]
GICF w/Embeddings	AR	0.882	[34]

Table 6. Datasets used in classification tasks.

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.

[58] Tijmen Tieleman. 2008. Training Restricted Boltzmann Machines Using Approximations to the Likelihood Gradient. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*. ACM, New York, NY, USA, 1064–1071.

https://doi.org/10.1145/1390156.1390290

[59] Amos Tversky. 1977. Features of similarity. *Psychological Review* 84, 4 (1977), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

[60] Laurens JP Van der Maaten, Eric O Postma, and H Jaap van den Herik. 2009. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research* 10, 1-41 (2009), 66–71.

[61] Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. (November 2016), 138–142. http://aclweb.org/anthology/W16-5618

[62] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. (June 2016), 88–93. http://www.aclweb.org/anthology/N16-2013

[63] Yunqing Xia, Erik Cambria, Amir Hussain, and Huan Zhao. 2015. Word Polarity Disambiguation Using Bayesian Model and Opinion-Level Features. *Cognitive Computation* 7, 3 (2015), 369–380. https://doi.org/10.1007/s12559-014-9298-4

[64] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A General Framework for Incorporating Knowledge into Word Representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 1219–1228. https://doi.org/10.1145/2661829.2662038

[65] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 673–681. https://doi.org/10.1145/3159652.3159703

[66] Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. 2016. A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition. *CoRR* abs/1606.06871 (2016). arXiv:1606.06871 http://arxiv.org/abs/1606.06871

[67] ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (March 2008), 137–213. https://doi.org/10.1561/1500000008

[68] ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (March 2008), 137–213. https://doi.org/10.1561/1500000008

[69] Ziqi Zhang and Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *CoRR* abs/1803.03662 (2018). arXiv:1803.03662 http://arxiv.org/abs/1803.03662

[70] Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-Adaptive Hierarchical Sentence Model. *CoRR* abs/1504.05070 (2015). arXiv:1504.05070 http://arxiv.org/abs/1504.05070