Transfer Learning: The Impact of Test Set Word Vectors, with Applications to Political Tweets

Nikhil Garg
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
nkgarg@stanford.edu

Arjun Seshadri
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
aseshadr@stanford.edu

Abstract

A major difficulty in applying deep learning in novel domains is the expense associated with acquiring sufficient training data. In this work, we extend literature in deep transfer learning by studying the role of initializing the embedding matrix with word vectors from GLoVe on a target dataset before training models with data from another domain. We study transfer learning on variants of four models (2 RNNs, a CNN, and an LSTM) and three datasets. We conclude that 1) the simple idea of initializing word vectors significantly and robustly improves transfer learning performance, 2) cross-domain learning occurs in fewer iterations than in-domain learning, considerably reduces train time, and 3) blending various out-of-domain datasets before training improves transfer learning. We then apply our models to a dataset of over 400k tweets by politicians, classifying sentiment and subjectivity vs. objectivity. This dataset was provided unlabelled, motivating an unsupervised and transfer learning approach. With transfer learning, we achieve reasonable performance on sentiment classification, but fail in classifying subjectivity vs. objectivity.

1 Introduction

One of the largest hurdles in applying deep learning to new domains is the lack of high quality training data. This is especially the case in domains in which computational methods, let alone deep learning, are rarely used. In this paper, we explore how to improve transfer learning, the process by which models trained in one domain are applied in a different domain. In particular, we focus on the role that initializing word vectors has in improving transfer learning performance. Though a controversial idea, we hypothesize and then conclude that initializing word vectors using the target test dataset significantly improves out-of-domain transfer learning.

Our particular context and initial motivation is analyzing a dataset of tweets by congresspeople over five years. Politicians often default to social media platforms such as Twitter to communicate with their constituents. Many congresspeople personally post on Twitter daily, expressing opinions, answering questions, and interacting with other politicians. This has resulted in a wealth of untapped data, perhaps the only such data, that could answer important questions about how congresspeople communicate. While tweets show great promise in revealing answers to these questions, the data is very sparse in sentiment content. Indeed, the vast majority of tweets from Representatives and Senators describe news or other emotionless aspects about themselves, while only a sparse selection reveals sentiment information. These facts directly undermine a hand labelling approach for training data, as sentiment laden tweets are sparse in the data. The approach is especially prohibitive for the state-of-the-art neural network models for sentiment classification, which often require more than 10k data points. These challenges are not unique to this particular scenario, but rather common to the general space of Twitter analysis, with the typical solution being to purchase sorted and labelled tweets, or financing a sizable labelling team.

Due to the constraints of our target dataset and interests in answering transfer learning questions, we take the following approach: First, we in an unsupervised manner sort the data into that which is likely to contain sentiment-rich information and that which is unlikely to do so. Then, we hand-label enough tweets (1300) for test datasets for sentiment and subjective vs. objective classification. Finally, we apply transfer learning from various datasets to train models for these tasks, paying particular attention to the role of word vectors in transfer learning performance. Across datasets from 3 different domains and 4 different deep learning models, we measure performance (mainly target dataset accuracy) with various word vector initializations across two tasks, sentiment analysis and subjective vs. objective classification. Our most robust result relates to this role, and we conclude that test dataset word vector initialization has a significant role to play in transfer learning.

2 Background & Related Work

There has been a recent growth in literature applying transfer learning with neural networks. A well known fact is that the hidden layers farther away from the output encode more fundamental information about a particular problem [1]. Today, the vast majority of deep learning problems in Computer Vision do not involve retraining a new neural network, but rather leverage this property by intelligently reusing networks already trained on large, well established datasets [2]. Transfer Learning has also had a growing presence in Deep NLP and Neural speech processing [3], with the most recent work focused on applying the methods to Machine Translation [4]. The specific problem of out-of-domain sentiment analysis discussed in this work has also been attempted [5, 6]. While the focus of that literature has been more about employing autoencoders to solve the problem, we use focus on the impact of different word vector initializations and dataset properties.

Political Science has yet to fully embrace the concepts behind machine learning, especially deep learning. The strong philosophical emphasis on avoiding ‘black box models’ makes references on many machine learning techniques scarce within the field. After an extensive literature review, we find only one work applying deep learning to political science [7]. While this dearth of literature in Political Science makes the applications of this work novel, the techniques, models, and technical analysis employed here are common to the entire deep learning community.

3 Approach

3.1 Data Pre-Processing

Prior to delving into the problem to any technical depth, we study the Political tweet dataset carefully through selective hand labelling. Here, we discover the sentiment sparsity of the dataset the vast majority of tweets by politicians are “objective” remarks of travel plans, news, and pleasantries that contain little sentiment content. Moreover, the dataset is riddled with tweets in foreign languages and messages that are gibberish, making hand labelling a multi-step process of first identifying the presence of sentiment, and then the sentiment expressed. This sparsity of sentiment in the Political dataset motivates the design of a semi-supervised pre-processing step to filter and isolate the desired data, as the alternative is hand labelling on the order of 100k tweets to retain roughly 10k sentiment filled tweets. We thus create a mass labelling scheme that separates tweets with subjective content from the tweets with objective content.

First, we use GloVe [8] to generate word vectors for the dataset. Then, we apply k-means to cluster the vectors formed by the average of the vectors of word vectors in each tweet. Because GloVe encodes words by their meaning and semantic structure, using k-means on means of these vectors allows for us to roughly cluster tweets with similar properties. Reading only a couple tweets from each cluster, we then label each cluster as containing wholly subjective, objective, or discarded content. Using 8 clusters results in 4 clusters of largely objective content, 2 of tweets with largely sentiment filled content, and 2 clusters consisting of tweets with foreign language and pleasantries that we could immediately discard. Our clustering mechanism has an 73.6% accuracy (as defined by labels found by individually examining tweets). Figure 1 shows our clusters with sample tweets.

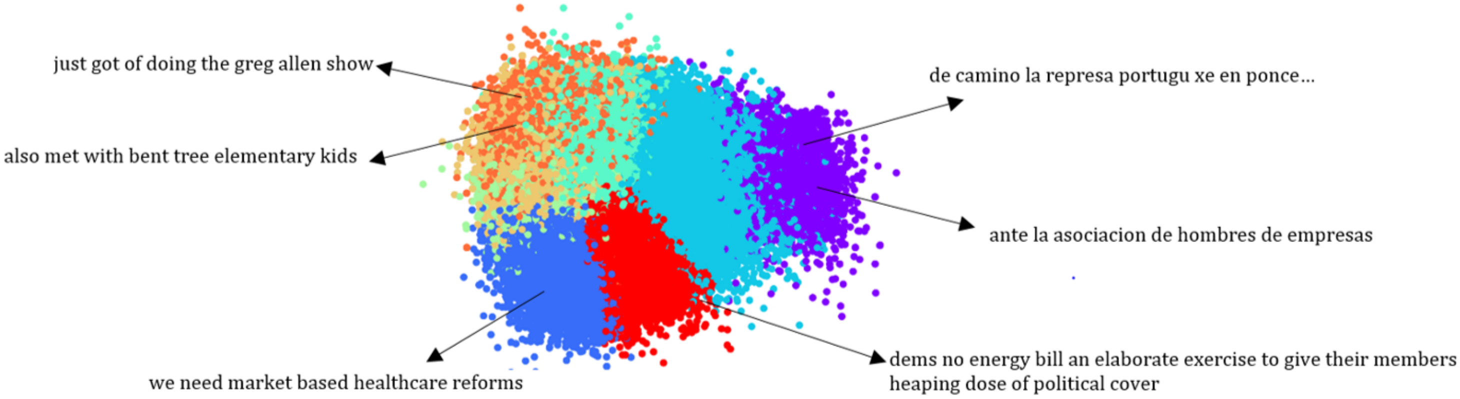


Figure 1: The 8 different clusters formed from the semi-supervised pre-processing projected to 2 dimensions, with sample tweets from the corresponding clusters.

3.2 Transfer Learning

The pre-processing step allows us to focus on a smaller subset of the data with much greater sentiment content. Directly hand labelling a training set is still somewhat intractable, as high performing neural network models often require well over 10k tweets to train to prevent over-fitting to the training set. Furthermore, because deep learning is new to Political Science, there exists no readily available pre-labelled dataset containing political language to serve as a proxy for the training set.

We thus leverage transfer learning as our main approach, training models in-domain where labelled data is plentiful and applying those models to our Political dataset. In particular, we analyze the effects that word vector initialization has on transfer learning performance. Transfer learning allows us to only hand label a substantially smaller validation set in the Political domain rather than having to label sizable training sets. We then leverage readily available sentiment datasets such as movie reviews and a general collection of tweets for our training data.

3.3 Neural Network Models

We are interested in the robustness of our results in transfer learning and thus build several models.

3.3.1 Deep Recurrent Neural Networks (RNN)

Our baseline model is a Deep RNN, chosen for its relative simplicity and high performance with short sentences like those found in Twitter data. Specifically, we use two variants of Deep RNNs. The first ("RNNWhole") contains the same number of time steps as the words in a tweet, while the second ("RNN"), a smaller, fixed number of time steps through which the words in a single tweet pass through batch by batch. To avoid the vanishing gradient problem with RNNWhole, the final hidden layer at every time step is summed together before the softmax step, ensuring that the gradient back-propagates to every time-step during training. Both models are structured to adjust their depth with the modification of single parameter, which allows us to treat model depth as a hyper-parameter to be tuned. In all of our experiments, the model depth is held below 7, further preventing any vanishing gradient effects.

3.3.2 Deep Long-Short-Term-Memories (LSTM) Networks

The LSTM is among the state-of-the-art neural networks in in-domain sentiment classification, making it a suitable choice for exploring out-of-domain performance. While training on a particular dataset, LSTM learns the memory structure associated with sentences of that dataset through the parameters that control its gates. Using an LSTM in an out-of-domain setting is then tantamount to asking the question of how memory structure in one domain transfers over to another, and whether one domain's memory structure is useful for sentiment classification in another. Similar to the two

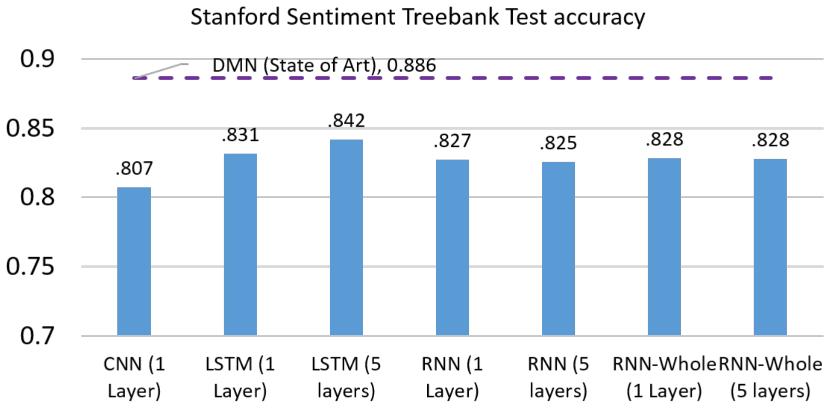


Figure 2: Accuracy of each model on the SST using standard train/test splits. Test accuracy is only calculated over root level sentences.

RNN models, the Deep LSTM is structured to scale to arbitrary depth, again making model depth a tunable hyper-parameter. As before, model depth is always held below 7.

3.3.3 Convolutional Neural Networks (CNN)

The CNN is yet another neural network with near state-of-the-art performance, making it a premier candidate for inclusion in our out-of-domain study. Because the CNN generalizes the recursive concepts found in a Recursive Neural Network, its performance can provide insight into how recursive structure transfers from one domain to another. While the CNN implemented in this work is not deep, it employs a variable number of filters and various filter types tunable as hyper-parameters. Please note that we did not implement the CNN from scratch but rather adapted an implementation available at `https://github.com/dennybritz/cnn-text-classification-tf`.

3.4 Benchmarks

The above networks form a useful basis to explore transfer learning in the context of Deep NLP. Prior to describing our exploration, we establish benchmark performances. We use the Stanford Sentiment Treebank (SST), a dataset with readily available performance metrics for the state-of-the-art networks [9]. Figure 2 displays the in-domain performance of our final, tuned networks as compared to the state-of-the-art. Note that the notion of performance here is somewhat misleading. The models presented in this work are tuned heavily for out-of-domain (transfer learning) performance, requiring vastly different hyper-parameters. It is interesting to see that despite these shortcomings, the networks still perform reasonably well compared to the current state-of-the-art, the Dynamic Memory Network [9].

3.5 Hand Labelling

While the application of Transfer Learning reduces the burden of hand labelling, a small, reliable set of labels is still necessary to validate the performance of the above models on the Political Dataset. We form this set by manually labelling over 1300 tweets, predominantly consisting of sentiment filled tweets, but also including objective content for comparison. Noteworthy here is the fact that the very concept of sentiment, or even the more fundamental notion of objectivity is by no means consistently held from one person to another. Thus, prior to the joint labelling effort, lengthy discussions are held about where the boundaries between these concepts lie when expressed in the form of tweets, eventually leading to a codebook that details rules and examples for consistent classification. A subset of the tweets is then labelled by both parties, and the resulting percentage of discrepancies are identified as the human error associated with the set.

4 Experiments

4.1 Transfer Learning

Our main focus is to develop methods for effective transfer learning between datasets in different domains, and we run extensive experiments across datasets, domains, and models to identify the most robust methods. After benchmarking our models on our datasets, we begin cross-domain training – training on one dataset and testing performance on another. We run each dataset combination several times, varying models and parameters. Our main hypothesis is that initializing word vectors on the test dataset during model training improves performance (and limiting the model vocabulary to the test dataset vocabulary has a similar but lesser effect), and so we train our models with various embedding matrix initializations.

4.2 Cross-Domain Sentiment, and Subjective vs. Objective, Classification

We take our best models (including model type, depth, and other hyper-parameters) and classify our political dataset, labelling each tweet as either subjective or objective. We further label each subjective tweet as having either negative or positive sentiment.

4.3 Datasets

Political Tweets Dataset of over 400k Tweets from Congress-people over five years, from which we hand-label 1300 tweets with Subjective/Objective and sentiment. 477/1300 (36.69%) of the tweets are Objective, leaving 823/1300 (63.31%) of the tweets as Subjective. Of the 823 Subjective tweets, 135 have neutral sentiment and are discarded. Of the 688 remaining tweets, 307 (44.6%) have positive sentiment and 371 (55.6%) have negative sentiment. These values are entirely determined through our labelling, which has 80.9% human agreement on Subjective/Objective and 93.8% agreement on binary sentiment labels.

Stanford Sentiment Treebank (SST) [10] We train on all root level sentences and a 33% subsample of phrases in the train split, and test on root level sentences in the test split.

Sentiment Analysis Dataset (SAD) [11] 1.5M tweets with sentiment labels (neg., neutral, pos.)

Movie Sentiment Dataset [12] 10k sentences with binary sentiment labels.

Movie Subjective/Objective Dataset [13] 10k labelled objective/subjective sentences; contains noticeable artifacts as objective synopses contain sentiment filled language.

4.4 Evaluation

We have several tools for evaluation. For out-of-domain classification, we measure performance (accuracy, precision, & recall) on a validation subset of our 'target' dataset. When comparing methods such as different word vector initializations, we primarily use accuracy. Furthermore, since we are interested in the training process itself for transfer learning, we observe train/in-domain-validation/target-domain-validation performance over time.

To the best of our knowledge, there are no prior benchmarks for deep transfer learning. However, we can evaluate the quality of our results through several mechanisms: first, by comparing our models to state-of-the-art for these tasks when both training and testing sets are drawn from the same dataset, as we present above in Section 3 (as discussed above, maximizing this performance is not our primary concern and, as discussed in Section 5.1, may be counter-intuitive); second, by comparing different methods (such as various word vector initializations) to each other across various settings and observing the robustness of the relationships. These comparisons allow us to analyze the effect of different methods even without benchmarking against the state-of-the-art in our specific task.

To evaluate our hypothesis that datasets that are more similar will perform better in transfer learning, we measure dataset similarity by the total variation distance : $\sum_{w\in\mathcal{X}\cup\mathcal{Y}}\frac{1}{2}|P_X(w) - P_Y(w)|$ where P_X, P_Y are the word probability distributions of the datasets, respectively, and $\mathcal{X}, \mathcal{Y}$ are their support. We report the relationship between the total variation distances and the best performance we could achieve with transfer learning, conditioned on each target dataset.

Finally, to evaluate performance on our specific tasks (political sentiment and subjective/objective classification), we compare our results to human 'accuracy' as defined as agreement during labelling.

5 Results

5.1 Transfer Learning

Our first class of results are in transfer learning. Figures 3, 4, and 5 show target dataset accuracies of each of our models after cross-domain training while initializing with various different word vectors. We note that across most of our datasets and models, initializing word vectors with GLoVe vectors from the test dataset increases performance, often significantly. Furthermore, when testing on the Political dataset, initializing word vectors using the training dataset (SST) performed no better than initializing with random word vectors and limiting the embedding matrix to either the test or train vocabulary. In other words, word vectors are very strong reflections of the dataset used to produce them and are not transferable.

Though we did not run this experiment due to time constraints, we expect these results to hold even when the training and testing datasets are drawn i.i.d. from the same corpus – initializing word vectors from the test dataset before training could achieve state-of-the-art performance with the right model (such as a Dynamic Memory Network).

Our next transfer learning result is displayed in the table inside Figure 6, which shows, for each target dataset, the accuracy achieved by using different training sets against the total variation distance between the datasets when using our CNN model. We find that, in general, the closer the two datasets in word probability distribution, the better the transfer performance. More significantly, we test the transfer learning performance when training on a mixture of out-of-domain training sets. Mixing datasets results both in a smaller total variation distance and a higher target set performance than any of the individual datasets that contribute to the mixture. This result can be made more robust by testing on more datasets and using different metrics than the total variation distance.

Finally, we find that when training on a different corpus, the model first learns general features that transfer across corpora and then corpus-specific features. Figure 7 shows accuracy and loss over time with our best model, LSTM with five layers. Even as the training dataset's train and validation set continue to rise, accuracy on the target dataset's validation set immediately peaks and starts falling. This result neatly extends intuition from known early stopping methods. Just as early stopping is used in in-domain training to prevent over-fitting to the train set, we find that aggressive early stopping over the target dataset's validation set must be used to prevent over-fitting to the training corpus. These results further suggest that tuning hyper-parameters to optimize in-domain validation performance is not optimal for out-of-domain performance.

Our performance with transfer learning on the SST does not reach the benchmarks presented in Section 3.4 for the SST, as expected, because there is a sharp transfer learning cost from using sets drawn from different corpora during test and training.

5.2 Sentiment Classification

Figure 3 shows our best accuracies for each model for binary sentiment classification on the political dataset using SST training, while Figure 7 includes precision and recall statistics for our best model, Deep LSTM. Our best model/dataset combination, with 78.5% accuracy, does not achieve close to human performance, with 93.86% accuracy. However, when taken in context to our best in-domain accuracy in the Stanford Sentiment Treebank (84.2%), the model performs reasonably considering noisy labels and out-of-domain training.

5.3 Objective vs. Subjective Classification

Figure 8 in the appendix shows our best accuracies for each model for binary Subjective/Objective classification on the political dataset using SST training, along with human performance and the performance of our initial semi-supervised clustering technique. In contrast to sentiment performance, the models did not perform well, barely beating random classification. This example represents a

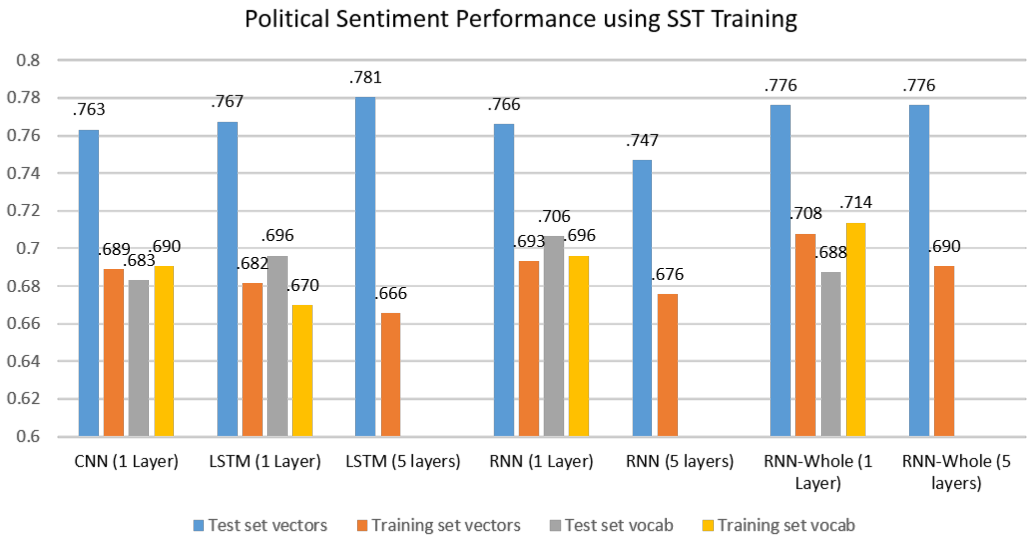


Figure 3: Accuracy of each model on the Political dataset while training on SST using different embedding matrix initializations and supports. Human agreement rate is 93.86%.

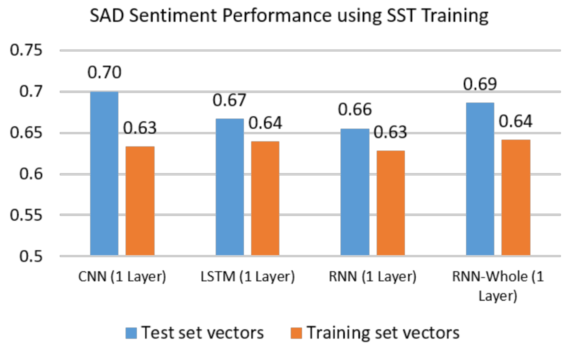


Figure 4: Accuracy of each model on the SAD dataset while training on SST using different embedding matrix initializations and supports.

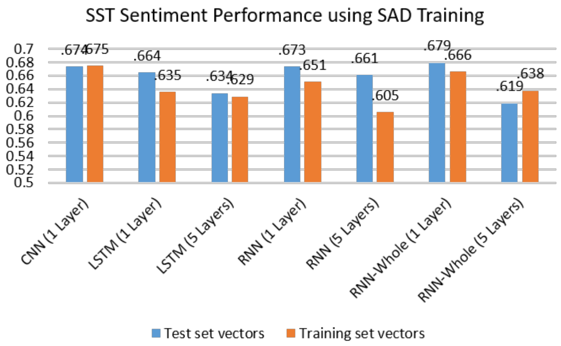


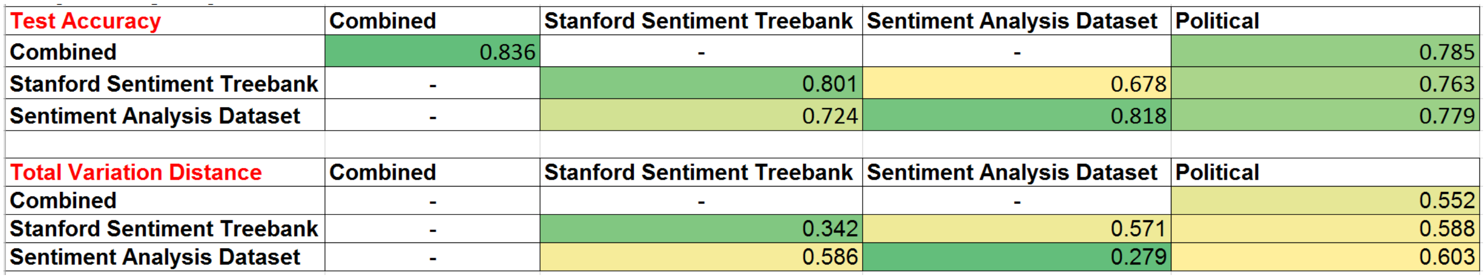
Figure 5: Accuracy of each model on the SST dataset while training on SAD using different embedding matrix initializations and supports.

failure in transfer learning because the models perform remarkably well in in-domain classification (near 90% in-domain test accuracy with several models).

6 Conclusion

In this report, we extend work in deep transfer learning by analyzing the role of word vector initialization, among other methods. We learn that using test dataset initialization for word vectors significantly improves performance in such settings, more so than any other parameter.

In future work, we suggest extending our results to standard in-domain training and evaluation – to study whether using word vectors initialized using the test set can improve test performance. Though such initialization would be difficult to implement in any online deep learning system, it could be used in applications where one wishes to analyze a large dataset in a single run.



Test Accuracy	Combined	Stanford Sentiment Treebank	Sentiment Analysis Dataset	Political
Combined	0.836	-	-	0.785
Stanford Sentiment Treebank	-	0.801	0.678	0.763
Sentiment Analysis Dataset	-	0.724	0.818	0.779

Total Variation Distance	Combined	Stanford Sentiment Treebank	Sentiment Analysis Dataset	Political
Combined	-	-	-	0.552
Stanford Sentiment Treebank	-	0.342	0.571	0.588
Sentiment Analysis Dataset	-	0.586	0.279	0.603

Figure 6: Target dataset accuracy and total variation distance, respectively, when training with the dataset on the left and testing on the dataset on the top, along with the total variation distances between the appropriate test/train splits. All numbers are using the single layer CNN model.

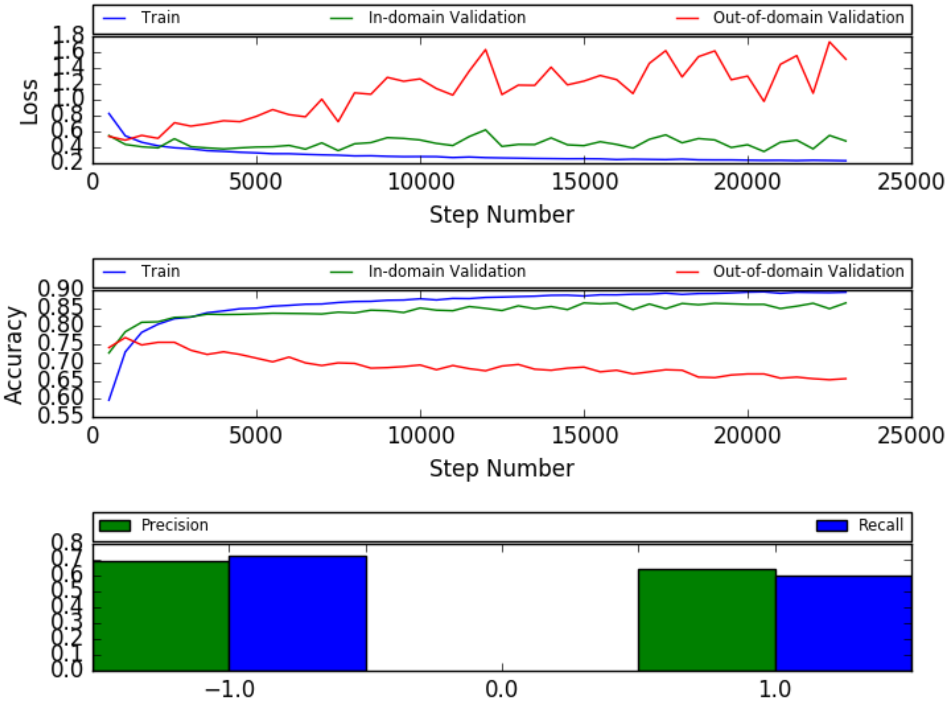


Figure 7: Loss and accuracy over time, with final precision/recall.

References

[1] Yann LeCun et al. "Object Recognition with Gradient-Based Learning". en. In: *Shape, Contour and Grouping in Computer Vision*. Lecture Notes in Computer Science 1681. DOI: 10.1007/3-540-46805-6_19. Springer Berlin Heidelberg, 1999, pp. 319–345. ISBN: 978-3-540-66722-3 978-3-540-46805-9. URL: http://link.springer.com/chapter/10.1007/3-540-46805-6_19 (visited on 06/04/2016).

[2] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: *Advances in Neural Information Processing Systems*. 2014, pp. 3320–3328. URL: http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks (visited on 06/02/2016).

[3] Dong Wang and Thomas Fang Zheng. "Transfer Learning for Speech and Language Processing". In: *arXiv:1511.06066 [cs]* (Nov. 2015). arXiv: 1511.06066. URL: http://arxiv.org/abs/1511.06066 (visited on 06/04/2016).

[4] Barret Zoph et al. "Transfer Learning for Low-Resource Neural Machine Translation". In: *arXiv:1604.02201 [cs]* (Apr. 2016). arXiv: 1604.02201. URL: http://arxiv.org/abs/1604.02201 (visited on 06/04/2016).

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *arXiv:1206.5538 [cs]* (June 2012). arXiv: 1206.5538. URL: http://arxiv.org/abs/1206.5538 (visited on 06/02/2016).

[6] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 513–520. URL: http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Glorot_342.pdf (visited on 06/04/2016).

[7] Won-Tae Joo, Y. S. Jeong, and KyoJoong Oh. "Political orientation detection on Korean newspapers via sentence embedding and deep learning". In: *2016 International Conference on Big Data and Smart Computing (BigComp)*. Jan. 2016, pp. 502–504. DOI: 10.1109/BIGCOMP.2016.7425979.

[8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[9] Ankit Kumar et al. "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing". In: *arXiv:1506.07285 [cs]* (June 2015). arXiv: 1506.07285. URL: http://arxiv.org/abs/1506.07285 (visited on 06/04/2016).

[10] Richard Socher et al. "Parsing With Compositional Vector Grammars". In: *In Proceedings of the ACL conference*. 2013.

[11] Links Naji. *Twitter Sentiment Analysis Training Corpus (Dataset)*. URL: http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/ (visited on 06/02/2016).

[12] Bo Pang and Lillian Lee. *Sentence Polarity Dataset v 1.0*. June 2004. URL: https://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz.

[13] Bo Pang and Lillian Lee. *Movie Review Subjectivity Dataset v 1.0*. June 2004. URL: http://www.cs.cornell.edu/people/pabo/movie-review-data/rotten_imdb.tar.gz.

Appendix

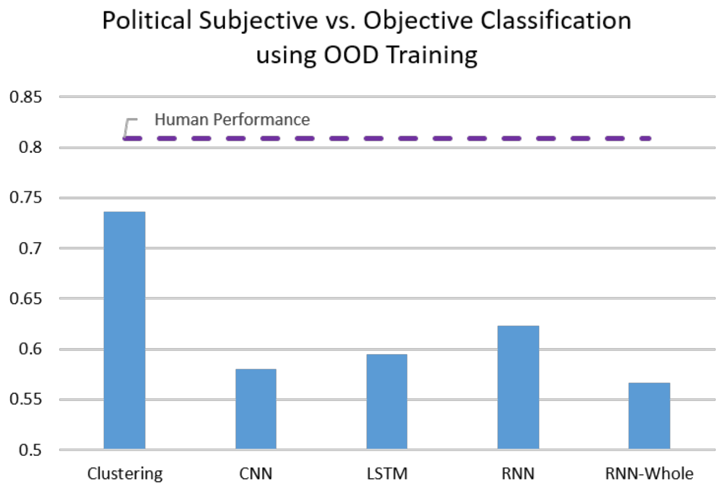


Figure 8: Test accuracy for Subjective/Objective Classification on the Political Dataset