A Shot in the Dark: Improved Zero-Shot and Few-Shot Transfer Learning with Self-Supervised Models for Sentiment Classification

Anwesha Mukherjee, Olivia Lee, Raj Palleti

Introduction & Motivation

The Small Data Problem:

- Prevalence of Large Language Models (LLMs) slowly increasing.
  - Often utilized for transfer learning.
- Sentiment Analysis and other tasks where data representation may be limited suffer from poor or negative transfer performance.
- Retraining is computationally expensive for big models.

Transfer Learning:

- Training a model for one task or data domain and adapting its use to another task or domain respectively.
  - **Domain Adaptation:** adapt with trained domain S (source) model applied to domain T (target) by constructing evaluation and test set from domain T.
    - **Zero-Shot:** Train - Domain S, Eval and Test - Domain T
    - **Few-Shot:** Train - Domain S + T_sample, Eval and Test - Domain T
- Successfully demonstrated on BERT, especially for cross-lingual transfer.
- Deep models struggle with up-to-par performance in transfer for sentiment analysis without similar corpuses in context.

Key Question: Are LLMs particularly better than ML methods and sequence models?

- Aim to model transfer learning with *self-supervised embeddings and supervised models* at various scales to optimize performance of sentiment classification compared to DistilBERT.
- Test direct tuning, zero-shot, and few-shot capabilities of these models and better understand respective limitations.

Data

Binary Sentiment Classification

Dead Poets Society has incredibly beautiful scriptwriting.	positive
365 Days is truly awful and deserved those 0% critic reviews	negative

Datasets

- 50K movie reviews from IMDb
- 9K tech product review tweets
- 50K uncategorized polar social tweets
- 50K movie reviews from Rotten Tomatoes

References

[1] R. Liu, Y. Shi, C. Ji and M. Jia, "A Survey of Sentiment Analysis Based on Transfer Learning," in IEEE Access, vol. 7, pp. 85401-85412, 2019, doi: 10.1109/ACCESS.2019.2925059.

[2] A. d. Arriba, M. Oriol and X. Franch, "Applying Transfer Learning to Sentiment Analysis in Social Media," 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), 2021, pp. 342-348, doi: 10.1109/REW53955.2021.00060.

[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.

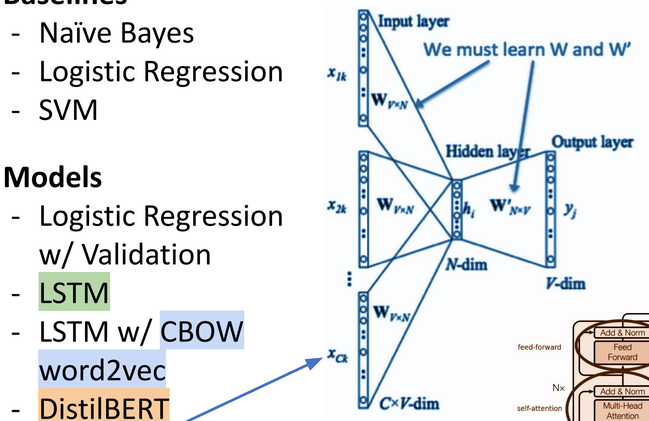
Selected Models and Architecture

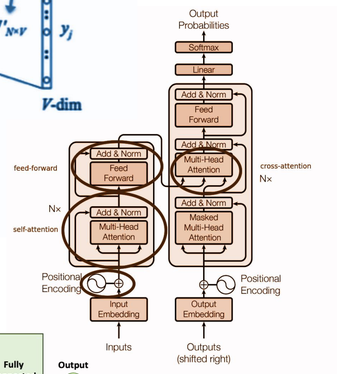
Baselines

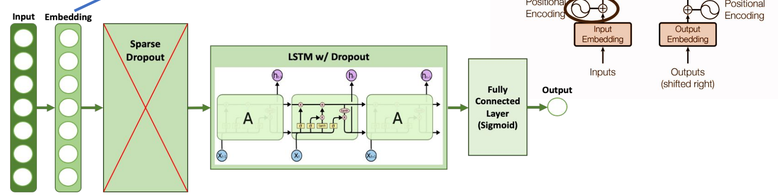
- Naïve Bayes
- Logistic Regression
- SVM

Models

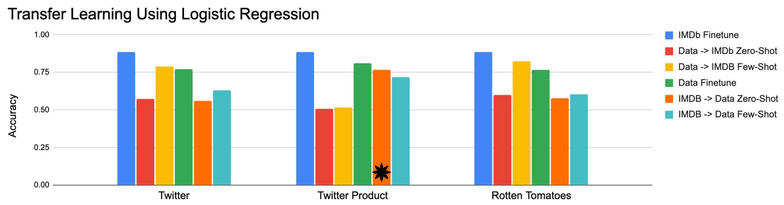
- Logistic Regression w/ Validation
- LSTM
- LSTM w/ CBOW word2vec
- DistilBERT

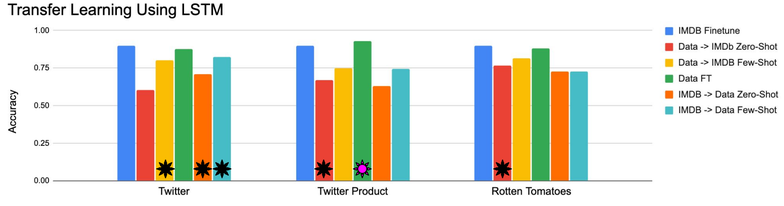


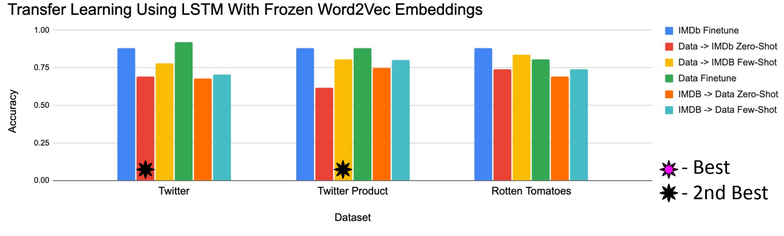




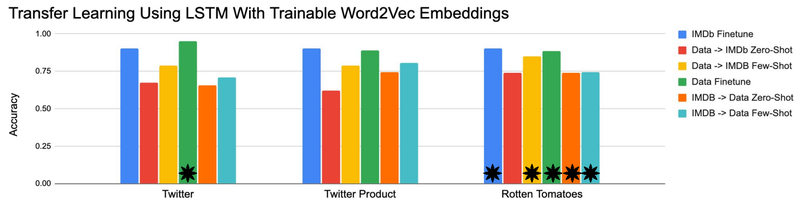
Experiment Results

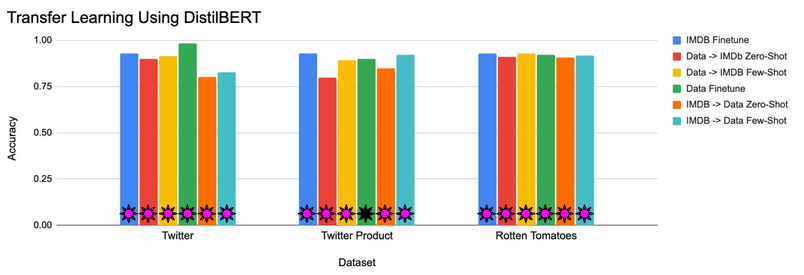






Experiment Results (cont'd) & Analysis





- **DistilBERT outperformed all** other **models** across the board
- Transferring from a larger dataset to a smaller dataset usually works better than the reverse direction.
- **Logistic regression** transfers much **better** to **smaller target data**
- **Trainable word2vec** worked particularly well between similar data domains (Rotten Tomatoes ↔ IMDb)

Conclusions & Future Work (Data Matters)

- Data polarity didn't play an important role opposed to **vocabulary variance and size**. Larger vocab better for general transfer.
- **Frozen** word2vec **embeddings** severely **limit model adaptability**
- **Zero-initialized** or **word2vec** trainable embeddings perform better in different contexts.
  - **Overfitting risk** when **data didn't overlap** using trainable word2vec.
- **Lack of validation gradient** hurts **log**istic **regression** performance.
- Pretrained attention schemes likely play large role in success of DistilBERT.
- Using trained embeddings from **word2vec** often reached convergence quicker than zero initialization regardless of performance
  - Immediate gradient impact.
  - Closer positioning to minimum loss

Future tests:

- Testing on rotten tomatoes with Frozen word2vec embeddings, zero-shot transfer improved by 4% in accuracy when augmenting the text8 (sampled Wikipedia data) corpus (**simulates pretrained embeddings**)
  - Testing this across datasets would be valuable.
- Testing **different activation functions** like gelu and relu on the LSTM to match DistilBERT and sharpen prediction boundaries.
- Stacking LSTMs to form more portable attention model.