The Use of Global Topic Information and Local Semantic Relationship in Sentiment Analysis

Jiachen Ning; Linhong Li; Shiqi Xiao; Xiangting Chen
Carnegie Mellon University

Overview

Background:

Sentiment analysis aims to extract subjective affect information from body of texts. The goal of this project is to build an interpretable sentiment analyzer on a dataset that contains a collection of 50,000 reviews posted on IMDB.

We explored effective sentiment analysis along two trajectories:

❖ ***Extracting representative text embeddings***

The most common text tokenization method is using N-grams. However, previous research argued that n-gram tokenization does not fully capture the gist of a document, and pairs of grammatically related words (i.e., dependency pairs) should also be incorporated as features as they might capture important long-distance information [3][4].

❖ ***Building an adaptive classification ensemble***

Neural networks perform most of the heavy-lifting in state-of-the-art pipelines, demonstrating the insufficiency of a single classifier like Naive Bayes or Support Vector Machine in performing the task. Topic modelling has been shown to improve sentiment classification accuracy on Twitter data [2].

Contributions:

Motivated by these two approaches, we proposed a feature vectorization scheme to capture grammatical dependency, and an ensemble of topic-specific classifiers to alleviate the burden of a single classifier having to learn complicated decision boundaries for the entire corpus.

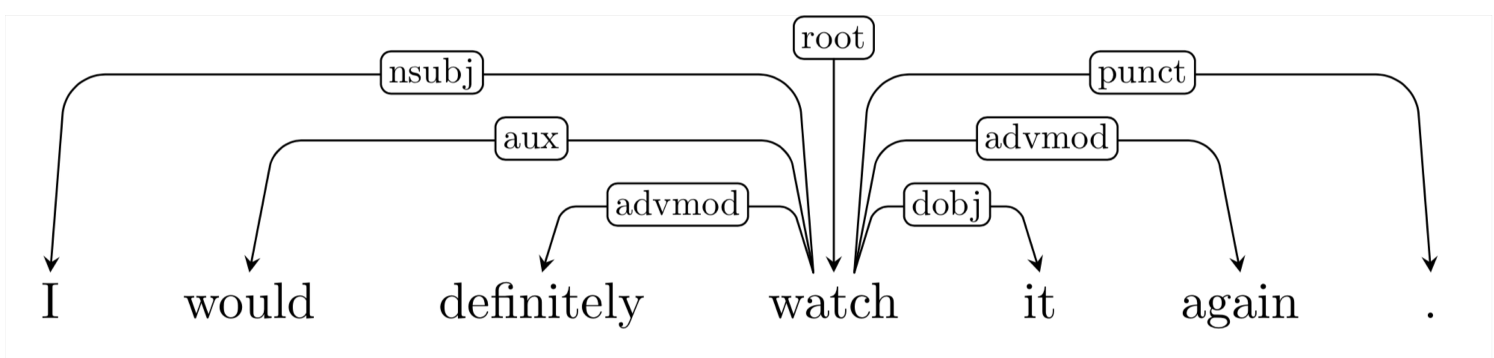

Figure 1. Example of Parser Output

Methods

Training:

1. Lemmatized the training data and eliminated HTML tags (e.g., '
').
2. Used Stanford Parser [1] to extract the lexical dependencies between words. Irrelevant dependencies were left out according to previous research [3].
3. Generated new bigrams from extracted dependency pairs. Combined with regular 1-3 gram tokens to create customized bag-of-words (BoW) vectors.
4. Modeled training set topic distributions by Latent Dirichlet Allocation (LDA).
5. Distributed texts to their most relevant classifiers for topic-specific training.

Testing:

1. Obtained testing set topic distributions using the trained topic model.
2. Passed testing data to all topic-specific classifiers. Weighted their predictions by topic relevance to get the final classifications.

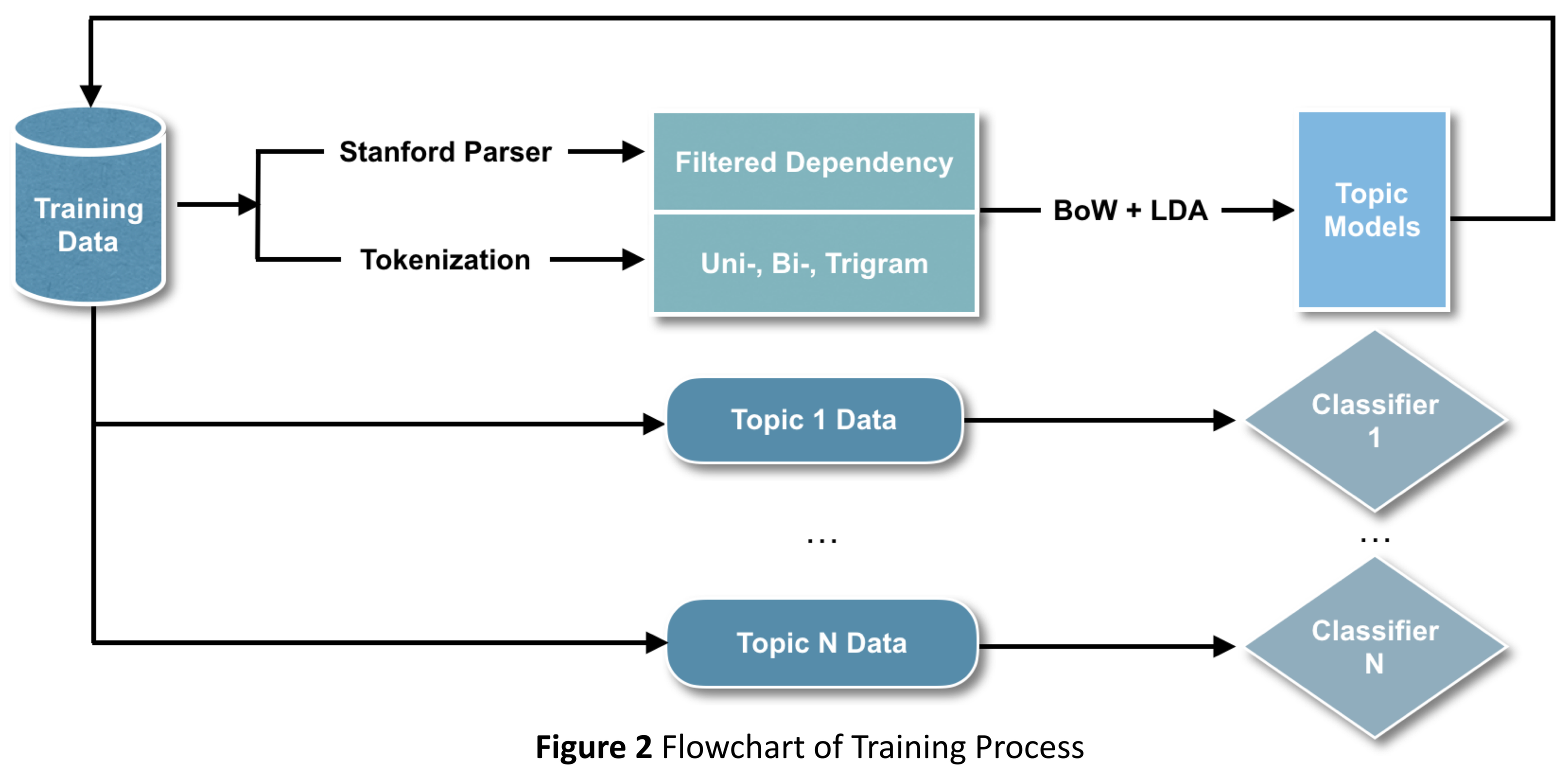

Figure 2 Flowchart of Training Process

Customized Bag-of-Words

The first group of experiments used a single classifier but with different feature vectorizations to test the efficacy of **customized BoW vectors**.

- Baseline accuracy obtained by using BoW vectorization and text-frequency (TF) vectorization on 1-3 gram tokenization.
- Lemmatization alone decreased classification accuracy.
- Combining lemmatized 1-3 gram tokens with dependency pairs in BoW vectors boosted test accuracy by **9.82%**.
- TF vectors improved test accuracy similarly (by 5.33%).

Token	Vector	Accuracy
1-3grams	BoW	0.84476
1-3grams	TF	0.85336
1-3grams + Lemmatize	BoW	0.83812
1-3grams + Lemmatize + Dependency	**BoW**	**0.94296**
1-3grams + Lemmatize + Dependency	TF	0.90668

Note: In all tests, we used NB.

Table 1 Experiments with Customized Bag-of-Words

Token & Vector	Topics	Accuracy
1-3grams & BoW	-	0.84792
1-3grams & BoW	**10**	**0.85972**
1-3grams & BoW	12	0.84792
1-3grams + Lemmatize + Dependency & BoW	**10**	**0.94296**

Note: In all tests, we used NB. In all tests with Topic Modeling, we used LDA and set the number of top topics to be 2.

Table 2 Experiments with Topic-Based Ensemble

Topic-based Ensemble

The second group of experiments only used regular BoW vectors to compare the performance of **a single classifier** and **a topic-based ensemble**.

During validation, topic-based ensemble outperformed a single classifier under specific hyperparameter settings. This indicates that the ***number of latent topics assumed*** affects how the topics are correlated with sentiments, and ***how biased the ensemble classifiers are*** affects their collective decisions. Table 2 reports test accuracy using the two best hyperparameter combinations.

Number of latent topics assumed:

- LDA hyperparameters were finetuned by 5-fold cross validation:
  - 10 latent topics yielded the lowest perplexity
  - 12 latent topics yielded the lowest validation error
- The number of latent topics that scored the lowest LDA perplexity achieved the best test accuracy and improved classification accuracy by **1.18%**.
- This suggests that a topic model that clusters the data well is suitable for building a topic-based ensemble that can enhance baseline performance.

How biased the ensemble classifiers are:

- Increasing the number of classifiers a single document is assigned to increases the number of examples each classifier sees, thus decreasing the topic-specific bias each classifier assumes.
- Figure 3 shows that higher classifier bias increased validation accuracy.

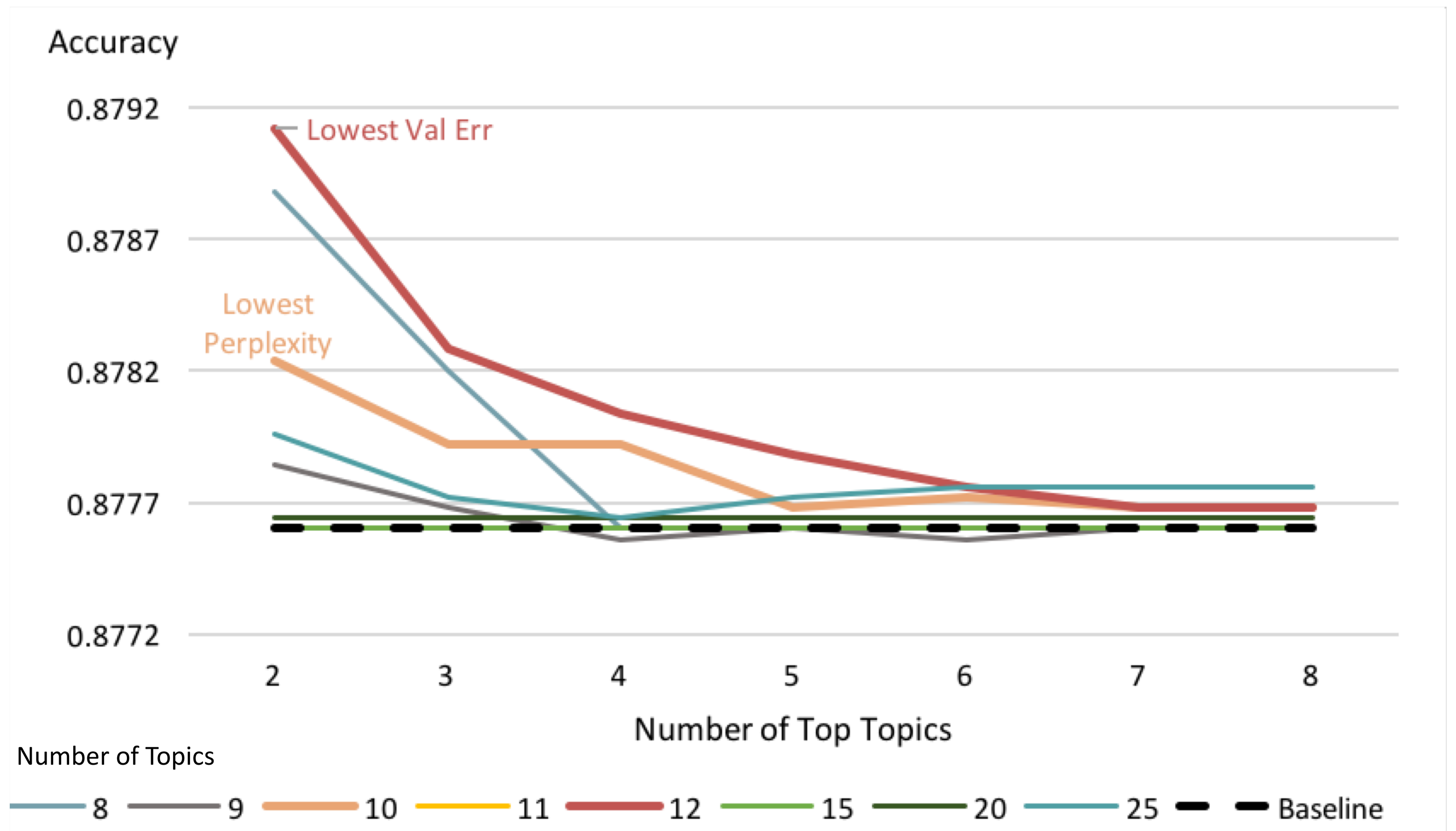

Figure 3 Validation Accuracy of Different Hyperparameter Settings

Future Work

- Combine customized feature vectors with the topic-specific classifiers, which requires fine-tuning the hyperparameters for the end-to-end experiments using cross-validation.
- Explore alternative vectorizations with rich information like lexical polarity.
- Analyze individual topic-specific classifiers for better model interpretability.

References

1. De Marneffe, M. C., & Manning, C. D. (2008). Stanford typed dependencies manual (pp. 338-345). Technical report, Stanford University.
2. Ficamos, P., & Liu, Y. (2016). A topic based approach for sentiment analysis on Twitter data. *International Journal of Advanced Computer Science and Applications*, *7*(12), 201-205.
3. Nastase, V., Shirabad, J. S., & Caropreso, M. F. (2006). Using dependency relations for text classification. In *Proceedings of the 19th Canadian conference on artificial intelligence* (pp. 12-25).
4. Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, *181*(6), 1138-1152.