Sentiment Analysis for Movie Reviews

Jiachen Ning Linhong Li Shiqi Xiao Xiangting Chen

March 6, 2019

1 Introduction

Sentiment Analysis is the task of classifying text documents by their *sentiment*, or the overall opinion towards the subject matter [4]. In practice, sentiment analysis is useful for companies and policymakers to monitor large scale public opinions on specific products and/or policies, and thus extended applications of sentiment analysis can be found in security, medical, finance, and entertainment domains [4].

The goal of this project is to build an interpretable sentiment classifier based not only on text content information, but also on lexical polarity information and unobserved topic information that could be derived from the text, using one of the most analyzed dataset for sentiment analysis – the IMDb Dataset of movie reviews. The dataset contains a collection of 50,000 reviews posted on the Internet Movie Database. Standard preprocessing ensured that the dataset is balanced, containing the same number of positive and negative reviews, where positive reviews had a score greater than or equal to 7 and negative reviews had a score less than or equal to 4 [3].

Our primary task is to output a binary classification of an audience's attitude towards a movie based on his or her online review. To do that, we will experiment with both content-based and topic-based input representations of the movie reviews on a variety of classification models. The performance of our model will be evaluated based on classification accuracy.

2 Literature Review

Literature review suggested that successful sentiment analysis requires finding (1) the most representative text embeddings and (2) the best performing classification scheme.

For the first requirement, it is a standard practice to tokenize the text into sequences of n words, called n-gram representations, with bigram and trigram representations being the most widely used and showing superior results than unigram tokens [7]. To vectorize a tokenized text, common methods include counting the token occurrences (i.e., Bag-of-Word, or BoW model) and assigning a term frequency (TF) or term frequency-inverse document frequency (TF-IDF) score to each token in the corpus dictionary, both of which reflect how important a token is to a document. However, the BoW and TF-IDF models are based solely on the text content. The joint sentiment/topic model proposed by Lin and He suggested that unsupervised topic modeling techniques such as LDA could potentially lead to more informative feature text representation for sentiment analysis [2]. Besides topic information, it was shown that weighting words by their lexical polarity and part of speech improves classification accuracy.

In terms of the second requirement for a successful sentiment analysis, several common choices of classifiers presented in previous works include Naïve Bayes (NB), Support Vector Machine(SVM), and (Deep) Neural Networks (NN) [1] [7] . For the IMDB dataset, the current best classifier is ULMFiT, which is a LSTM-based model proposed by Howard and Ruder in 2018 [1]; SVM with NB features (NBSVM) proposed by Wang and Mannings showed superior performance among non-NN classifiers [7].

3 Methods

Currently, most existing works mainly focused on testing various classifiers on content-based feature vectors. Thus, we explored different combinations of feature vectorization (BoW, TF, and TF-IDF) and the two most effective classifiers used in sentiment analysis, NB and SVM. For the SVM model, we used both linear and RBF kernels. Note that we fixed the tokenization to be unigram and bigram since it tends to give better performance [5]. All baseline models are constructed, trained, and tested using the machine learning package, scikit-learn [7].

4 Preliminary Results

Table 1: Baseline Models and Test Accuracy

Tokenization	Vectorization	SVM-Linear	SVM-RBF	NB
Unigram + Bigram	BoW	0.83212	0.63148	0.84272
Unigram + Bigram	TF	**0.8756**	**0.674**	0.85372
Unigram + Bigram	TF-IDF	0.8742	0.65448	**0.85476**

The combination of different vectorization methods and classifiers, as well as their respective testing prediction accuracies was given in Table 1.

5 Evaluation of Preliminary Work

In terms of classifiers, we can see that when using a BoW model, the best test accuracy was given by NB; when using either TF or TF-IDF as feature vectors, the best test accuracy was achieved by linear SVM. The reason why SVM that using RBF kernel performed poorly is because we didn't perform cross-validation during training to select the optimal gamma parameter and used the default value given by the package instead.

In terms of feature vectorization methods, we observed that both TF and TF-IDF give similar results and outperformed BoW model.

6 Future Work

7 Teammates and Work Division

References

[1] Howard, J., Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

[2] Lin, C., He, Y. (2009, November). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 375-384). ACM.

[3] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 142-150). Association for Computational Linguistics.

[4] Mäntylä, M. V., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis?A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16-32.

[5] Pang, B., Lee, L., Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

[6] Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research 12* (pp. 2825-2830).

[7] Wang, S., Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 90-94). Association for Computational Linguistics.