

ENSEMBLE OF GENERATIVE AND DISCRIMINATIVE TECHNIQUES FOR SENTIMENT ANALYSIS OF MOVIE REVIEWS

Grégoire Mesnil

University of Montréal

University of Rouen

Tomas Mikolov & Marc'Aurelio Ranzato

Facebook Artificial Intelligence Research

Yoshua Bengio

University of Montréal

ABSTRACT

Sentiment analysis is a common task in natural language processing that aims to detect polarity of a text document (typically a consumer review). In the simplest settings, we discriminate only between positive and negative sentiment, turning the task into a standard binary classification problem. We compare several machine learning approaches to this problem, and combine them to achieve a new state of the art. We show how to use for this task the standard generative language models, which are slightly complementary to the state of the art techniques. We achieve strong results on a well-known dataset of IMDB movie reviews. Our results are easily reproducible, as we publish also the code needed to repeat the experiments. This should simplify further advance of the state of the art, as other researchers can combine their techniques with ours with little effort.

1 INTRODUCTION

Sentiment analysis is among the most popular, simple and useful tasks in natural language processing. It aims at predicting the attitude of text, typically a sentence or a review. For instance, movies or restaurant are often rated with a certain number of stars, which indicate the degree to which the reviewer was satisfied.

This task is often considered as one of the simplest in NLP because basic machine learning techniques can yield strong baselines (Wang & Manning, 2012), often beating much more intricate approaches (Socher et al., 2011). In the simplest settings, this task can be seen as a binary classification between positive and negative sentiment.

However, there are several challenges towards achieving the best possible accuracy. It is not obvious how to represent variable length documents beyond simple bag of words approaches that lose word order information. One can use advanced machine learning techniques such as recurrent neural networks and their variations (Mikolov et al., 2010; Socher et al., 2011), however it is not clear if these provide any significant gain over simple bag-of-words and bag-of-ngram techniques (Pang & Lee, 2008; Wang & Manning, 2012).

In this work, we compared several different approaches and realized, without much surprise, that model combination performs better than any individual technique. The ensemble best benefits from models that are complementary, thus having diverse set of techniques is desirable. The vast majority of models proposed in the literature are discriminative in nature, as their parameters are tuned for the classification task directly. In this work, we boost the performance of the ensemble by considering a generative language model. To this end, we train two language models, one on the positive reviews and one on the negative ones, and use the likelihood ratio of these two models

evaluated on the test data as an additional feature. For example, we assume that a positive review will have higher likelihood to be generated by a model that was trained on a large set of positive reviews, and lower likelihood given the negative model. In this paper, we constrained our work to binary classification where we trained two generative models, positive and negative. One could consider a higher number of classes since this approach scales linearly with the number of models to be train, i.e. one for each class. The large pool of diverse models is a) simple to implement (in line with previous work by Wang and Manning (Wang & Manning, 2012)) and b) it yields state of the art performance on one of the largest publicly available benchmarks of movie reviews, the Stanford IMDB dataset of reviews. Code to reproduce our experiments is available at <https://github.com/mesnilgr/iclr15>.

2 DESCRIPTION OF THE MODELS

In this section we describe in detail the approaches we considered in our study. The novelty of this paper consists in combining both generative and discriminative models together for sentiment prediction.

2.1 GENERATIVE MODEL

A generative model defines a distribution over the input. By training a generative model for each class, we can then use Bayes rule to predict which class a test sample belongs to. More formally, given a dataset of pairs $\{x^{(i)}, y^{(i)}\}_{i=1, \dots, N}$ where $x^{(i)}$ is the i -th document in the training set, $y^{(i)} \in \{-1, +1\}$ is the corresponding label and N is the number of training samples, we train two models: $p^+(x|y = +1)$ for $\{x^{(i)} \text{ subject to } y^{(i)} = +1\}$ and $p^-(x|y = -1)$ for $\{x \text{ subject to } y = -1\}$. Then, given an input x at test time we compute the ratio (derived from Bayes rule): $r = p^+(x|y = +1)/p^-(x|y = -1) \times p(y = +1)/p(y = -1)$. If $r > 1$, then x is assigned to the positive class, otherwise to the negative class.

We have a few different choices of distribution we can choose from. The most common one is the n -gram, a count-based non-parametric method to compute $p(x_k^{(i)} | x_{k-1}^{(i)}, x_{k-2}^{(i)}, \dots, x_{k-N+1}^{(i)})$, where $x_k^{(i)}$ is the k -th word in the i -th document. In order to compute the likelihood of a document, we use the Markov assumption and simply multiply the n -gram probabilities over all words in the document: $p(x^{(i)}) = \prod_{k=1}^K p(x_k^{(i)} | x_{k-1}^{(i)}, x_{k-2}^{(i)}, \dots, x_{k-N+1}^{(i)})$. As mentioned before, we train one n -gram language model using the positive documents and one model using the negative ones.

In our experiments, we used SRILM toolkit (Stolcke et al., 2002) to train the n -gram language models using modified Kneser-Ney smoothing (Kneser & Ney, 1995). Furthermore, as both language models are trained on different datasets, there is a mismatch between vocabularies: some words can appear only in one of the training sets. This can be a problem during scoring, as the test data contain novel words that were not seen in at least one of the training datasets. To avoid this problem, it is needed to add penalty during scoring for each out of vocabulary word.

N -grams are a very simple data-driven way to build language models. However, they suffer from both data sparsity and large memory requirement. Since the number of word combinations grows exponentially with the length of the context, there is always little data to accurately estimate probabilities for higher order n -grams.

In contrast with N -grams languages models, Recurrent neural networks (RNNs) (Mikolov et al., 2010) are parametric models that can address these issues. The inner architecture of the RNNs gives them potentially infinite context window, allowing them to perform smoother predictions. We know that in practice, the context window is limited due to exploding and vanishing gradients (Pascanu et al., 2012). Still, RNNs outperform significantly n -grams and are the state of the art for statistical language modeling. A review of these techniques is beyond the scope of this short paper and we point the reader to (Mikolov, 2012) for a more in depth discussion on this topic.

Both when using n -grams and RNNs, we compute the probability of the test document belonging to the positive and negative class via Bayes' rule. These scores are then averaged in the ensemble with other models, as explained in Section 2.4.

Table 1: Performance of SVM with Wang & Manning (2012) rescaling for different N-grams

| Input features | Accuracy |
|---------------------------|----------|
| Unigrams | 88.61% |
| Unigrams+Bigrams | 91.56% |
| Unigrams+Bigrams+Trigrams | 91.87% |

2.2 LINEAR CLASSIFICATION OF WEIGHTED N-GRAM FEATURES

Among purely discriminative methods, the most popular choice is a linear classifier on top of a bag-of-word representation of the document. The input representation is usually a tf-idf weighted word counts of the document. In order to preserve local ordering of the words, a better representation would consider also the position-independent n-gram counts of the document (bag-of-n-grams).

In our ensemble, we used a supervised reweighing of the counts as in the Naive Bayes Support Vector Machine (NB-SVM) approach (Wang & Manning, 2012). This approach computes a log-ratio vector between the average word counts extracted from positive documents and the average word counts extracted from negative documents. The input to the logistic regression classifier corresponds to the log-ratio vector multiplied by the binary pattern for each word in the document vector. Note that the logistic regression can be replaced by a linear SVM. Our implementation¹ slightly improved the performance reported in (Wang & Manning, 2012) by adding tri-grams (improvement of +0.6%), as shown in Table 1.

2.3 SENTENCE VECTORS

Recently, (Le & Mikolov, 2014) proposed an unsupervised method to learn distributed representations of words and paragraphs. The key idea is to learn a compact representation of a word or paragraph by predicting nearby words in a fixed context window. This captures co-occurrence statistics and it learns embeddings of words and paragraphs that capture rich semantics. Synonym words and similar paragraphs often are surrounded by similar context, and therefore, they will be mapped into nearby feature vectors (and vice versa).

Such embeddings can then be used to represent a new document (for instance, by averaging the representations of the paragraphs that constitute the document) via a fixed size feature vector. The authors then use such a document descriptor as input to a one hidden layer neural network for sentiment discrimination.

2.4 MODEL ENSEMBLE

In this work, we combine the log probability scores of the above mentioned models via linear interpolation. More formally, we define the overall probability score as the weighted geometric mean of baseline models: $p(y = +1|x) = \prod p^k(y = +1|x)^{\alpha_k}$, with $\alpha_k > 0$.

We find the best setting of weights via brute force grid search, quantizing the coefficient values in the interval $[0, 1]$ at increments of 0.1. The search is evaluated on a validation set to avoid overfitting. We do not focus on a smarter way to find the α since we consider only 3 models in our approach and we consider it out of the scope of this paper. Using more models would make the use of such method prohibitive. For a larger number of models, one might want to consider random search of the α coefficients or even Bayesian approaches as these techniques will give better running time performance.

3 RESULTS

In this section we report results on one of the largest publicly available sentiment analysis datasets, the IMDB dataset of movie reviews. The dataset consists of 50,000 movie reviews which are categorized as being either positive or negative. We use 25,000 reviews for training and the rest for

¹<https://github.com/mesnilgr/nbsvm>

Table 2: Performance of Individual Models

| Single Methods | Accuracy |
|------------------|----------|
| N-gram | 86.5% |
| RNN-LM | 86.6% |
| Sentence Vectors | 89.3% |
| NB-SVM Trigram | 91.87% |

Table 3: Performance of Different Model Combinations

| Ensemble | Accuracy |
|------------------------------------|---------------|
| RNN-LM + NB SVM Trigram | 92.13% |
| RNN-LM + Sentence Vectors | 90.4% |
| Sentence Vectors + NB-SVM Trigrams | 92.39% |
| All | 92.57% |
| State of the art | 91.22% |

testing, using the same protocol proposed by (Maas et al., 2011). All experiments can be reproduced using the code available at <https://github.com/mesnilgr/iclr15>.

Table 2 reports the results of each individual model. We have found that generative models performed the worst, with RNNs slightly better than n-grams. The most competitive method is the method based on reweighed bag-of-words (Wang & Manning, 2012)². Favoring simplicity and reproducibility of our performance, all results reported in this paper were produced by a linear classifier.

Finally, Table 3 reports the results of combining the previous models into an ensemble. When we interpolate the scores of RNN, sentence vectors and NB-SVM, we achieve a new state-of-the-art performance of 92.57%, to be compared to 91.22% reported by (Wang & Manning, 2012). Notice that our implementation of the Sentence Vectors method (Le & Mikolov, 2014) alone yielded only 89.3% (a difference of $\simeq 3\%$). In order to measure the contribution of each model to the final ensemble classifier, we remove one model at a time from the ensemble. We observe that the removal of the generative model affects the least the ensemble performance. Overall, all three models contribute to the success of the overall ensemble, suggesting that these three models pick up complimentary features useful for discrimination. In Table 4, we show test reviews misclassified by single models but classified accurately by the ensemble.

4 CONCLUSION

We have proposed a very simple yet powerful ensemble system for sentiment analysis. We combine three rather complementary and conceptually different baseline models: one based on a generative approach (language models), one based on continuous representations of sentences and one based on a clever reweighing of tf-idf bag-of-word representation of the document. Each such model contributes to the success of the overall system, achieving the new state of the art performance on the challenging IMDB movie review dataset. Code to reproduce our experiments is available at: <https://github.com/mesnilgr/iclr15>. We hope researchers will take advantage of our code to include their new results into our ensemble and focus on improving the state of the art for Sentiment Analysis.

REFERENCES

Kneser, Reinhard and Ney, Hermann. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*,

²In our experiments, to match the results from (Le & Mikolov, 2014), we followed the suggestion by Quoc Le to use hierarchical softmax instead of negative sampling. However, this produces the 92.6% accuracy result only when the training and test data are not shuffled. Thus, we consider this result to be invalid.

Table 4: Reviews misclassified by Single Models but classified accurately by the Ensemble

| Model | Sentences |
|-----------------|--|
| NB-SVM | <p>(positive) a really realistic , sensible movie by ramgopal verma . no stupidity like songs as in other hindi movies . class acting by nana patekar . much similarities to real 'encounters' .</p> <p>(negative) leslie nielson is a very talented actor , who made a huge mistake by doing this film . it doesn't even come close to being funny . the best word to describe it is stupid !</p> |
| RNN-LM | <p>(positive) this is a good film . this is very funny . yet after this film there were no good ernest films !</p> <p>(negative) a real hoot , unintentionally . sidney portier's character is so sweet and lovable you want to smack him . nothing about this movie rings true . and it's boring to boot .</p> |
| Sentence Vector | <p>(positive) this movie is based on the novel island of dr . moreau by version by john frankenheimer .</p> <p>(negative) if it wasn't for the terrific music , i would not hesitate to give this cinematic underachievement 2/10 . but the music actually makes me like certain passages , and so i give it 5/10 .</p> |

volume 1, pp. 181–184. IEEE, 1995.

Le, Quoc V. and Mikolov, Tomas. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 2014.

Maas, Andrew L., Daly, Raymond E., Pham, Peter T., Huang, Dan, Ng, Andrew Y., and Potts, Christopher. Learning word vectors for sentiment analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2011.

Mikolov, Tomáš. *Statistical language models based on neural networks*. PhD thesis, 2012.

Mikolov, Tomas, Karafiát, Martin, Burget, Lukas, Cernocký, Jan, and Khudanpur, Sanjeev. Recurrent neural network based language model. In *INTERSPEECH*, pp. 1045–1048, 2010.

Pang, Bo and Lee, Lillian. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Pascanu, Razvan, Mikolov, Tomas, and Bengio, Yoshua. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*, 2012.

Socher, Richard, Pennington, Jeffrey, Huang, Eric, Ng, Andrew, and Manning, Christopher D. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Conference on Empirical Methods in Natural Language Processing*, 2011.

Stolcke, Andreas et al. Srlm-an extensible language modeling toolkit. In *INTERSPEECH*, 2002.

Wang, Sida and Manning, Christopher D. 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, 2012.