Restaulytics: Sentiment Analysis in Restaurant Reviews

Georgiana Denisa Stefan
Faculty of Automatic Control and Computers
University Politehnica of Bucharest
Bucharest, Romania
georgiana.stefan94@stud.acs.upb.ro

ABSTRACT

Research is quintessential to the marketing decisions of an organization, as it offers the means to study the consumer's perceptions and behavior. Nowadays, the process of gathering the data for this research can be not only "push-based", but also "pull-based", as users can freely and publicly express their opinions in the online medium through reviews. However, it is hard to make sense of such a large quantity of unstructured information. But this is where natural language processing and machine learning come in hand. Since we are interested in the sentiment uncovered by the linguistic patterns as an expression of customer perception, we built a model that can automatically determine the number of stars consumers would rate a restaurant based on the reviews they leave, so as to offer marketers a quantitative way of assessing performance. For that, we turned to deep learning and convolutional neural networks trained on what we call "restaurant sentiment embeddings". Designing our own sentiment embeddings instead of using the pretrained GloVe ones has already improved the network's performance by 3%, reaching 66% state-of-the-art accuracy on a 5-classes classification. However, we aim to further increase this performance by adding handcrafted features expressing topic information extracted via Latent Dirichlet Allocation, as well as the sentiment associated with each topic. While the idea of learning domain specific sentiment embeddings is not new, our approach is different from existing ones [1] in the input and model used for learning. Moreover, when feeding a convolutional network with these sentiment embeddings, we report higher accuracy than other similar convolutional approaches [2,3,4].

KEYWORDS

Sentiment Analysis, Convolutional Neural Network, Latent Dirichlet Allocation, GloVe

1 Methodology

1.1 Dataset

In order to train and evaluate the model, we used a subset of 3.2 million restaurant reviews from the Yelp! Open Dataset (https://www.yelp.com/dataset).

1.2 Features

1.2.1 Restaurant sentiment embeddings. Word embeddings provide us with a way of representing words in a computer-intelligible way that is still able to preserve some meaning through the concept of distance. Two words that are close in this continuous space determined by the embeddings are also similar in concept. We chose GloVe pre-trained word vectors (https://nlp.stanford.edu/projects/glove/) as starting point for learning and then fine-tuned them on the restaurant dataset, as it is known that words can change meaning when used in different contexts. In order to do that, we used a neural network that would take as input the review in embedded form, with embeddings trainable, and fed them to a fully-connected softmax layer that would output for a review the probability of belonging to each of the 5 classes. The embeddings were learned by optimizing the cross-entropy function and saved for future use as features.

1.2.2 Handcrafted features. Latent Dirichlet Allocation models a review as a distribution of topics and a topic as a distribution of words. It is an unsupervised learning approach that aims to discover the grouping of words into topics, starting from given initial distributions and number of topics. Through manual examination and perplexity and UMass comparison of models with different numbers of topics, we reached the conclusion that the 25-topic model is the most appropriate for the task at hand. Assuming that the topics presented in a review have also an impact on the final review score, we aim to include the review representation as topic distribution in the set of input features, which is a novel approach. Moreover, for the most prominent topics in a review, we aim to compute a sentiment score based on the positive and negative children in the dependency graph of its most important words.

1.3 Model

The sentiment embeddings are fed to a convolutional neural network with the following architecture:

- an input layer taking in the review as a vector of 1234 word embeddings
- a convolutional layer with a filter size of 7, stride 1 and ReLu as the activation function - the filter size was determined experimentally; as it turns out, 7 is the number that best leverages closeness and context
- a max pooling layer extracting the most prominent features in the activation maps
- a fully-connected softmax layer

The model was trained using 80% of the data and evaluated on the remaining 20%. The training stopped after 3 epochs in which accuracy did not improve on the validation set. In the end, it reached 66.03% accuracy on the test set.

2 Conclusions and Future work

We present a simple, yet efficient way of learning domain specific sentiment embeddings that can be further used to improve the accuracy of convolutional neural networks, reaching state-of-the-art results. In this way, we validate the hypothesis that knowing the context of words, both by using domain specific embeddings and convolution filters, is important when decoding a message. Future work will be conducted to improve this accuracy by adding the handcrafted features to the model and fine-tune it.

REFERENCES

[1] D. Tang et al., 2014. Learning sentiment-specific word embedding for twitther sentiment classification. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1.

[2] C. Dos Santos and M. Gatti, 2014. Deep convolutional neural networks for sentiment analysis on short texts. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers

[3] N. Kalchbrenner, E. Grefenstette and P. Blunsom, 2014. A convolutional neural network for modelling sentences.

[4] S. Lai et al., 2015. Recurrent convolutional neural networks for text classification. *Twenty-ninth AAAI conference on artificial intelligence.*