Yelp Text Reviews Rating Prediction

Ashley Levato, Yi Jin, Umber Singh

University of California Berkeley

December 16, 2016

Abstract

The paper describes the work experimenting with applications of natural language processing (NLP) techniques and convolutional neural networks (CNN) to increase the accuracy of predicting star ratings from yelp user reviews. The CNN approach performs better than optimized naive bayes techniques, and enables analysis of predictor words using the CNN parameters instead of standard statistics. Continued experimentation can further optimize the CNN approach with parameter turning and addressing class imbalance issues.

I. Introduction

THIS paper describes work using NLP techniques to predict a particular review's Yelp star rating from the associated text. The work will be focused on the available Yelp challenge dataset comprised of 2.7M reviews from 687K users for 85K businesses.

While convolutional neural networks (CNN) are widely used in image recognition, many suggest that the importance of invariance to shifts and rotations in objects of images loses value with NLP, a domain in which the order of words can be instrumental, emphasizing the way in which humans process words. This notion leads to the use of Long Short-Term Memory and other recurrent neural networks (RNN) as the preferred NLP technique since it sequences and maintains an internal representation of words one at a time.

However, reviews can be comprised of smaller ideas and phrases. Leveraging CNNs in the context of Yelp's text reviews are better suited with a model that weighs word and character patterns seen repeatedly pieced together to create meaning instead of the fixed "memory" extracting long-range context. The CNN approach extracts meaning from sentiment scale/polarity analysis.

II. Background

The Yelp dataset and challenge are constantly evolving. The data set for this project is pulled from challenge round 8. [1] Previous work using this dataset has investigated rating predictions using regression, support vector machine, and naive bayes models. [2, 3] In some studies these models are also being used to create recommendations for users based on their reviews. This work focuses on using NLP techniques and specifically CNNs to build upon the previous work and attempt to identify a more accurate method.

Just from the name, Neural Networks can be said to be a machine learning methodology which is figured out and modeled with a very close reference to the structure of the brain. It is made up of neurons which is a network of learning units which convert input signals, for instance the picture of cat, into corresponding output signals, for instance the "cat" label, basically achieving automated recognition [6]. Taking an example of automatic image recognition, the whole process of ascertaining that a picture contains a cat much involves a function of activation. If the picture copycats prior images of a cat the neurons have come across before, the "cat" label will be automatically activated. This means that the more the images the neurons have had an exposure to, the better its learning on recognizing other images which are not labeled, which is called the process of neurons training, an uncanny process [6]. Convolutional Neural Network aka CNN is a feed-forward artificial neural network type whereby the pattern of connectivity between its neurons is much inspired by the way animal visual cortex is organized making it a biologically-inspired model. Convolutional Neural Network is a type of artificial neural network that differs from the normal typical neural networks by the way in which signal flow between neurons. In the regular neural networks signals are passed in a single direction along the input-output channel, without allowing for a feedback or loopback of signals into the network aka in a forward feed technique only [6]. Through forward feed networks can be successfully used in text and image recognition, the technique prerequisites the connection of all neurons thus an network structure which is overly-complex consequently constituting to a high complexity cost and very long training times when large data sets are used in training the network. Due to this shortcomings in the typical neural networks, there was dire need for better networks, hence the Convolutional Neural Networks [4].

CNNs work by making predictions through learning the relationship there is between the features of a certain data and some priory observed response. In CNNs, which are more of filtering and encoding by transformation, each network layer works as a detection filter for the availability of some specified features or patterns that are present in the original data. The very first layers of the CNN recognize the large features which can be detected and interpreted easily with the later layers used to detect smaller and more abstract features that are usually present in many of the large features priory detected by the first layers. The last layer of the Convolutional Neural Network is responsible for making a classification that is ultra-specific through bringing together all the specific features that were detected by the previous layers in the input data [5].

In CNNs, each and every network layer works as a filter for the availability of specific patterns and features that are present in the original image. For detection by such an ultra-specific filter the position of the original image is irrelevant and the filters are made in such a way to detect whether or not the image contains such patterns or features. In the CNN learning process, the filter is positional shifted many times and applied at many different positions such that the entire image is covered fully in detail. Mostly, the input data is translated to image data first [5].

CNNs are very useful in natural programming processing. In the problem solving and identification of contexts as done by CNNs, it is able to correct morphological information and misspelled words or context identification. CNNs is useful in Natural Programming Processing where semantics similarity is important to be deduced with thelgorithm working even better for unseen words. Word and character embedding is very possible with CNNs in Natural Programming Processing and the identification of word sequence and association is very possible with CNNs. Slot filling and sequence labeling problems since the introduction of CNNs in Natural Programming Language has brought much accuracy [5].

The distinct advantage of CNNs is the accuracy in the recognition of images solving problems that have been faced before in concern with image recognition. CNNs come with lots of accuracy and supports large input image size with good computational cost and representation power and less loss of spatial power. However, CNNs are very complicated and need a lot of training on data management thus constituting high operational and computational costs. CNNs are so much complex to implement from scratch with machine practitioners often utilizing complex functions like Tensorflow, MatConvNet, Torch and Caffe [6].

CNNs are applied in day to day living and tackles broad spectrum of classification and prediction problems and in scaling up applications which would have taken intractable data amounts. The cellular networks that are everyday thing over the world is applying the CNNs allowing parameter sharing and convolution with CNN algorithms analogous for the use of flashcards in studying for exams. The gradient descent algorithms sustain forward learning for sequential flashcards randomly [5]. Learning algorithms require feedback for effective communication requires feedback. Through the validation of CNNs comparisons and predictions are easy to make with minimal errors through the back propagation of errors and requires much of CNNs. CNNs are so effective and have taken technology and programming to a higher ground giving good results that might not have been achieved otherwise.

In conclusion, CNNs is made of four distinct layers which include: Convolution, Subsampling, Activation and Fully Connected. Convolution is made up of the very first layers that are responsible for receiving input signals aka convolution filters. In convolution, the network labels the input signal in reference to the past learned originality. Subsampling includes the smoothening of inputs in the convolution layer to reduce the sensitivity of filters to variations and noise. Activation layer is responsible for controlling how the signal flow across and in each layer just like the way neurons are fired in the brain [4]. The Fully Connected layers include the last layers of the CNNs with neurons of the leading layers connected to each and every neuron in subsequent layers resembling the high reasoning levels where every possible pathway from the input to the output is considered.

III. Methods

This work addresses applications of CNNs to the Yelp Review dataset. Although, the dataset is in a semi-structured form, it still required a significant amount of data cleaning. The review text was analyzed and cleaned, by stop words removal, word stemming, finding the complete vocabulary, then padded out to all be consistent length for the CNN input, and translated into a numeric vector representation of the words. In some iterations of the model, additional parameter tweaking was performed including techniques of applying 0.005% document frequency filter to reduce the vocabulary size from 480K to about 23K.

Some previous work started with the Word2vec embedding. However, this project leverages the CNN and tensorflow to create a custom embedding for this dataset in order to associate the words with review characteristics instead of like words.[7] For instance, in a Word2vec scenario, all of the numbers are mapped to nearby vector space. So the phrases ”2 stars” and ”5 stars” would be mapped close together. However for predicting star ratings the two phrases are associated with very different review ratings. By allowing the NN to train the embeddings, the associations are specific to those for Yelp reviews and not general language use.

Some experiments involved different tuning of the model parameters.

- A large embedding size reduces the training speed much, but the accuracy is even a little worse since the word length is limited. The final model has embedding size set to 32.
- The number of filters reflects the number of neurons for each filter size. A lower filter size results poor accuracy. The final model has the number of filters set to 128.
- Considering the word segments, the filter width in the final model is 3, 4, and 5.

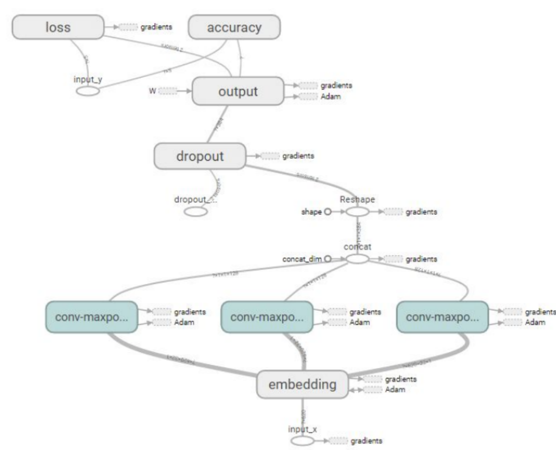


Figure 1: The CNN Model Structure

IV. Discussion

The baseline is bag of words model and multinomial NB, providing a 0.568 accuracy and 0.564 F1 score. The CNN model returns .62 accuracy and .618 F1 score. The metrics keeps improving when the training continues. But compared with bag of words model, the training process is very time consuming. The bag of words model is still efficient and successful in sentiment analysis.

Table 1: *Accuracy Comparison*

Model	Accuracy	F1 score
Bag of words	0.56	0.564
CNN (v3)	0.558	0.518
CNN (v4)	0.62	0.618

As shown in Figure 2, the accuracy continually increased with additional training (batch size 500): The training accuracy is a little higher than that on dev and test sets. One possible reason is the review texts changes among different industries. Shuttling the whole dataset may provide a better result.

A confusion matrix was generated based on the evaluation of a test set and the results are shown in Figure 3. The density of misclassified reviews are all within 1 degree of the diagonal indicating that there are no large misclassifications occurring. The confusion matrix highlights the most misclassified group is labeling some 4 star reviews as 5 stars, which could be the result of the class imbalance. As seen in Figure 1, the number of 5 star reviews is much greater than other ratings which could lead to biasing in predicting a rating of 5 instead of 4. This highlights a potential area for further development in the model.

Additional analysis focused on the strongest predictor words from each rating level of reviews. For every word in the vocabulary, the single word was run through the CNN as a single word review rating. These predicted star probabilities for each star value were used as a proxy for the word frequency in the word cloud. Word clouds of for each review level are shown in Figure 4 (on the last page). Many of the words make intuitive sense, but there are some interesting occurrences that emerged. For instance, in the 4 star reviews several of the prominent words including ”docked”, ”loses”, ”downfall” traditionally carry a negative connotation. However, they end up being strong predictors for a 4 star review because they appear in phrases like ”docking 1 star”, ”loses 1

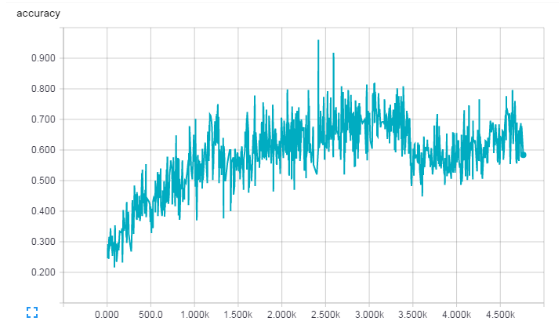


Figure 2(a): The accuracy during CNN training

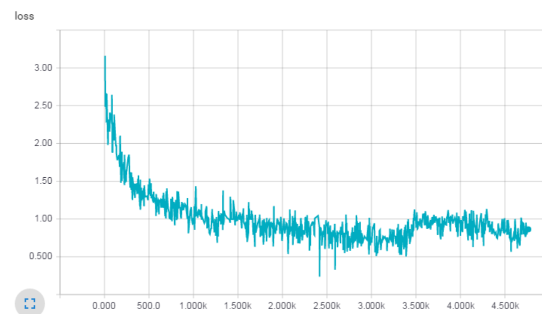


Figure 2(b): The loss during CNN training

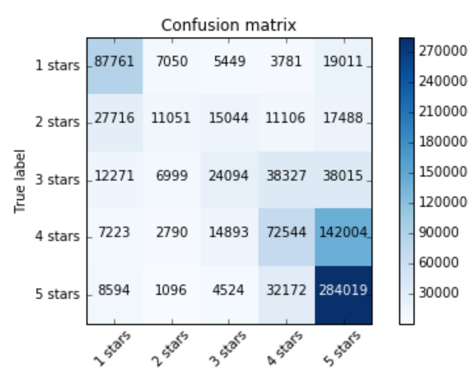


Figure 3: The confusion matrix

star", "the only downfall". In the 5 star word cloud some terms also appear that are indicative of certain service industries. For instance the word "locksmith" appears in the bottom right, this could indicate some services tend to have higher review ratings than other sectors.

Overall, utilizing NLP techniques and implementing the CNN improved the accuracy of rating predictions, but additional model tuning could continually improve the accuracy. Further work looking at particular locations or sectors of services may also be beneficial to improve accuracy of the model.

V. References

[1] Yelp Dataset Challenge `https://www.Yelp.com/dataset_challenge`

[2] Channapragada, S. & Shivaswamy R. Prediction of Rating Based on Review Text of Yelp Reviews `http://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/031.pdf`

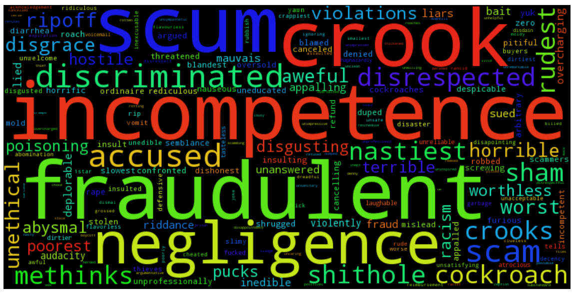
[3] Li, C & Zhang, J. Prediction of Yelp review Star Rating using Sentiment Analysis. CS229 Project Final report. `http://cs229.stanford.edu/proj2014/Chen%20Li,%20Jin%20Zhang,%20Prediction%20of%20Yelp%20Review%20Star%20Rating%20using%20Sentiment%20Analysis.pdf`

[4] Venugopal, K.R. & Patnaik, L.M. (2011). Computer Networks and Intelligent Computing: 5th International Conference on Information Processing, ICIP 2011, Bangalore, India, August 5-7, 2011. Proceedings,Springer Science & Business Media.

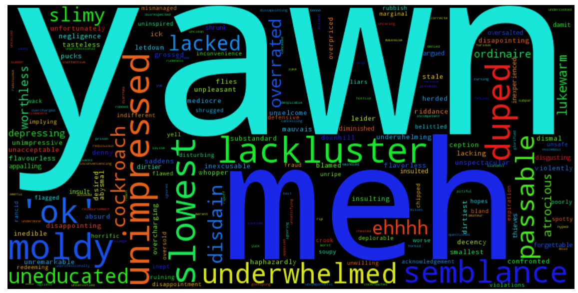
[5] Jayne, C. & Iliadis, L. (2016). Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings, Springer.

[6] Sharp, B. & Delmonte, R. (2015). Natural Language Processing and Cognitive Science: Proceedings 2014, Walter de Gruyter GmbH & Co KG.

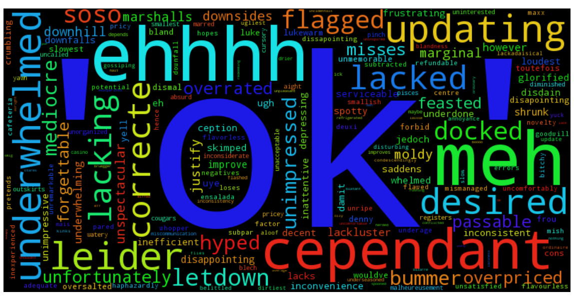
[7] WILDML, Implementing a CNN for Text Classification in Tensorflow.`http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/`



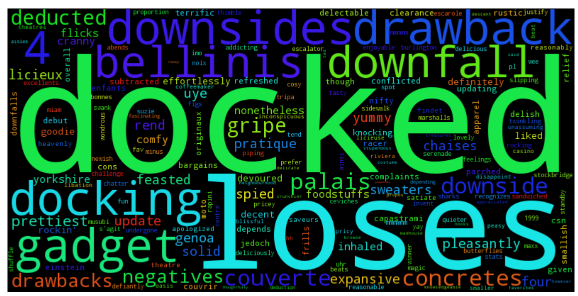
(a)



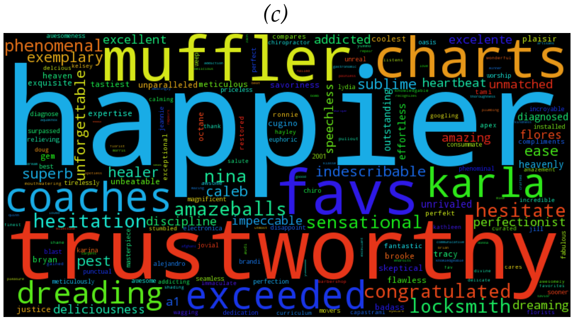
(b)



(c)



(d)



(e)

Figure 4: Each word cloud represents the best predictor words for each star rating level (a)1 star (b) 2 stars (c) 3 stars (d) 4 stars (e) 5 stars