Sentiment Analysis of Amazon Reviews using Neural Networks, Logistic Regression, and Pott's Scores

by Charles Dickstein

Abstract

My research experimented with sentiment analysis using Neural Networks, logistic regression, and naive bayes based systems to predict the ratings of Amazon product reviews. Using over 500 thousand Amazon reviews, three techniques for sentiment analysis were developed to predict the rating of an Amazon review based on text. Of the three systems a Dense Neural Network achieved the highest score to predict reviews by category while a Pott's Diagram performed the worst.

Introduction

On many internet forums a large amount of posts and comments referring to products are unrated. However, on sites such as Amazon users provide detailed reviews with star ratings on a scale from 1 to 5. These rated reviews serve as a viable source of labeled data set for sentiment analysis, which aim to classify a body of text based on semantic orientation towards a specific subject or on a general positive or negative basis. Since 58% of people report that online data is missing, impossible to find, or confusing and about 73% of people say reviews have a large influence on online purchases, it is plausible that tools that can do sentiment analysis of unrated reviews could have utility to both consumers and businesses [1].

The two most common methods for sentiment analysis are text classification and lexicon analysis. The lexicon approach typically uses dictionaries of words with prerated sentiment values such as positive or negative . A body of text is then analyzed and compared to the dictionary; the idea being that the sum total of words by their dictionary sentiment score will provide a reasonable prediction of the body of text's overall sentiment. Meanwhile, text classification usually involves using statistical algorithms to make predictions of a text's sentiment [2].

Sentiment Analysis Using Neural Networks

Whereas logistic regression outputs to one layer, a neural network outputs to a series of layers providing a more robust analysis of data. The input data is fed into a N-hidden layers, which contains a series of vectors known as nodes. Depending on the type of neural network the nodes values may be calculated differently, but in the case of a recurrent neural network used for my research, the formula is roughly as follows:

$$\boldsymbol{h}^{(t)} = f(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}; \boldsymbol{\theta}),$$

Any given node is a function of the current node and its preceding node, which means that the current node is the multiplicative of all prior nodes visited [3].

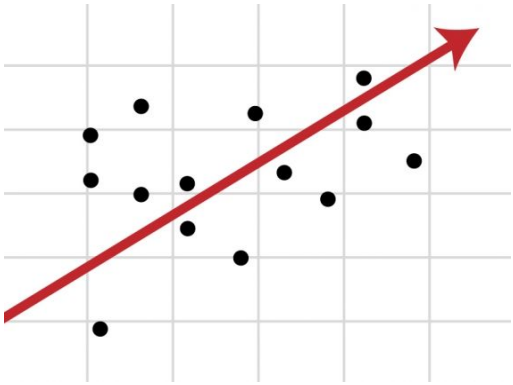
Additionally, neural networks utilize what is known as an activation layer, which allows for a highly accurate modelling of nonlinear data. Whereas other models of prediction such as naive bayes and logistic regression are linear, a neural network is often able to more accurately

[1] Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/150000001

[2] Lexicon-Based Methods for Sentiment Analysis Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede Computational Linguistics 2011 37:2, 267-307

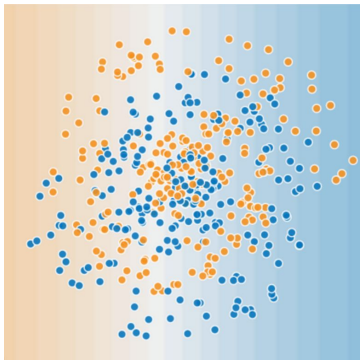
[3] Ian Goodfellow, Yoshua Bengio & Aaron Courville (2016). *Deep Learning.* MIT Press

model data that does not follow a linear pattern[4]. For example, the following picture shows a linear graph of an arbitrary x and y value. These values are therefore related in a linear manner that only requires a linear modelling system.



5

However, the second graph shown below shows a graph where the data is not related linearly, but nonlinearly. The creation of a linear model to predict this data would be very difficult[6].



7

Using a dense neural network I experimented with activation layers with Relu and sigmoid activation layers. Relu performed the best achieving an f-score of 62% and an accuracy of 67%. Sigmoid performed the worst of the two with an f-score of 54%.

Dataset

The dataset of Amazon reviewed was sourced from a csv file containing over 500 thousand Amazon product reviews found on Kaggle. Each review contained several labels such as a title, text, score, product id, helpfulness and many more. The labels used to build models for the algorithms were just the text and score. Moreover, each system separated the data into three sets for training, development, and testing. Each algorithm was evaluated by accuracy and f-score on the test set.

A known issue with data is its potential to have an uneven distribution of labels. Naive bayes and linear regression are especially susceptible to this issue as compared to neural networks due to its reliance on linear models. The most used method to fix this skew is through a method known as binning, which involves filtering a data set to create a more even dispersion of categories so as to eliminate a bias towards any given label[8]. The original dataset employed contained about 500 thousand reviews, however, over 70 percent of those reviews were rated 4 or 5 stars. Therefore, a bias existed for 4 and 5 star rated reviews. In order to account for this bias, a form of binning was used known as equal-width binning where the data was divided into an equal

---

[4] Mengqiu Wang & Christopher D. Manning (). *Effect of Non-linear Deep Architecture in Sequence Labeling.*

[5] Dizikes, Peter. "Explained: Regression Analysis." *MIT News*, 16 Mar. 2010, news.mit.edu/2010/explained-reg-analysis-0316.

[6] "Introduction to Neural Networks: Anatomy | Machine Learning Crash Course | Google Developers." *Google*, Google, 10 Apr. 2018,

[7] Developers." *Google*, Google, 10 Apr. 2018, developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/anatomy.

[8] Zeng, Guoping. A necessary condition for a good binning algorithm in credit scoring Applied Mathematical Sciences, Vol. 8, 2014, no. 65, 3229-3242 http://dx.doi.org/10.12988/ams.2014.44300

amount of each category winnowing the dataset down to 150 thousand reviews of 30 thousand reviews in each star category.

Pott's Diagram

Many forms of lexicon sentiment analysis exist focused on predefined sentiment. The method of lexicon based sentiment analysis studied was a modified naive bayes algorithm called a Pott's Diagram [9]. Naive bayes is an algorithm that can be used for text classification using an unordered set of data. Generally, naive bayes takes a function of the likelihood of a word in a category and the probability that a given word precedes or succeeds a current word. The Pott's Diagram is a supervised algorithm based on naive bayes, but modified specifically for sentiment [10].

A Pott's diagram is determined by its frequency across an individual category divided by the frequency of that word across all categories.

$$P(w|c) = \frac{count(w,c)}{\sum_c P(w|c)}$$

$$Pott's\ score = \frac{P(w|c)}{\sum_c P(w|c)}$$

The Pott's Score represents the overall probability a word is in a category. Once every word in a review's category has been collected, the system takes the multiplicative of each word's Pott's Score and the highest score by category is chosen as the review's prediction. In that sense, Pott's Score has a similar approach to both sentiment and lexicon analysis.

Results of Sentiment Analysis Using Neural Networks

The neural network system predicted Amazon reviews by modelling a dense neural network, which converts all of the words in a review to a vector as the input layer. Each vector is compared to other vectors in the same layer, which is then compared to vectors in other layers. Backpropagation occurs such that data in each neuron is fed backward to allow for temporal analysis. While naive bayes is naive due to its bag of words approach, a dense network network is able to look at a higher degree of connectivity between layers while still taking in as input a bag of words vector[11]. Using a Dense Neural network API's from Tensorflow, the system achieved a 62% f-score and an accuracy of 66% ratings from a scale of 1 to 5 proving to be the most reliable algorithm of all the techniques researched.

The two biggest components that affected the neural networks were epochs and the activation layer used. Each additional epoch allowed for a better model yet also became more computationally expensive. Using a ReLu activation layer, an epoch of five obtained a 60% f-score. Meanwhile, an epoch of eight obtained a 62% f-score. Meanwhile, an eight epoch network using a sigmoid activation layer obtained an f-score of 54% , which was eight percentage points worse than a Relu activation layer of 62%. In short, ReLu proved to be the most effective activation layer.

Results of Logistic Regression

[9] Jurafsky D, Martin JH (2008) Speech and language processing, prentice hall series in artificial intelligence. 2. Prentice Hall, NJ

[10] Jurafsky D, Martin JH (2008) Speech and language processing, prentice hall series in artificial intelligence. 2. Prentice Hall, NJ

[11] Chris Manning, Richard Socher,. "Deep Learning for NLP (without Magic)" Stanford University

With a model based on logistic regression the system obtained an f-score of 91% and an accuracy of 86%. In short, logistic regression works very well for predicting general sentiment.

Results of Pott's Diagram

The original biased data set obtained a 77% accuracy of general sentiment and a 62% accuracy on a scale from 1 to 5. However, due to the exorbitant bias towards five and four star reviews it is likely the original data was overfitting a large amount of data. However, after binning the data, the Pott's Diagram proved almost entirely ineffective.

Its f-score hovered at about 20% for predicting categories and 50% for predicting general sentiment. In other words, the algorithm was about as good as randomly guessing categories. Therefore, either not enough data was available for the algorithm to properly function or the algorithm may simply not be effective for diverse categories of sentiment based text classification.

Results of Deep Learning

The system predicted Amazon reviews by modelling a dense neural network, which converts all of the words in a review to a vector as the input layer. Each vector is compared to other vectors in the same layer, which is then compared to vectors in other layers. Backpropagation occurs such that data in each neuron is fed backward to allow for temporal analysis. While naive bayes is naive due to its bag of words approach, a dense network network is able to look at a higher degree of connectivity between layers while still taking in as input a bag of words vector. Using a Dense Neural network api from Keras, the system achieved a 66% accuracy of predicting ratings from a scale of 1 to 5 proving to be the most reliable algorithm of all the techniques researched.

Potential False Positives In Amazon Reviews

Although Amazon product reviews are great for sentiment analysis since they use a clear labelling system, users sometimes write contradicting text with their review. For example, a user may give a one star rating yet state that they loved the product. However, I had to make the assumption that the majority of ratings reflect the actual sentiment of a rating.

Comparison to other Research

The best f-score I attained for categories was a 0.62 using a dense neural network and LSTM neural networks. In comparison to other sentiment analysis systems, the system performed very well. In the SemEval-2017 Task 4 where a competition was held for modelling twitter data, a 0.62 f-score was in the top 10 out of forty teams with 0.67 being number one [12].

Additionally, my logistic regression model performed well for labelling positive and negative statements at an 0.91 f-score, tj.

Conclusion

Amazon Product Reviews are a useful source for sentiment analysis due to their clearly labeled ratings. Neural networks are exceedingly useful for sentiment analysis across multiple classes, while logistic regression works great for predicting general sentiment. Meanwhile.while a Pott's Diagram is ineffective at least with sparse

[12] Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pages 312–320, Atlanta, US

data. Due to the success of neural networks in this study, it is plausible that neural network based systems will continue to be the amongst the best method of both sentiment analysis and text classification across other text driven highly correlative features.

References

Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends® in Information Retrieval: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/150000001

Chris Manning, Richard Socher,. "Deep Learning for NLP (without Magic)" Stanford University

Dizikes, Peter. "Explained: Regression Analysis." *MIT News*, 16 Mar. 2010, news.mit.edu/2010/explained-reg-analysis-0316.

Guoping Zeng A necessary condition for a good binning algorithm in credit scoring . Applied Mathematical Sciences, Vol. 8, 2014, no. 65, 3229-3242. http://dx.doi.org/10.12988/ams.2014.44300

Ian Goodfellow, Yoshua Bengio & Aaron Courville (2016). *Deep Learning.* MIT Press

Jurafsky D, Martin JH (2008) Speech and language processing, prentice hall series in artificial intelligence. 2. Prentice Hall, NJGoogle Scholar

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-Based Methods for Sentiment Analysis**.** Computational Linguistics 2011 37:2, 267-307

Mengqiu Wang & Christopher D. Manning (). *Effect of Non-linear Deep Architecture in Sequence Labeling.*

Pintu Lohar, Koel Dutta Chowdhury, Haithem Afli, Mohammed Hasanuzzaman and Andy Way. ADAPT at IJCNLP-2017 Task 4: A Multinomial Naive Bayes Classification Approach for Customer Feedback Analysis task. In the Proceedings of IJCNLP, Shared Tasks, pp. 161–169, 2017. AFNLP.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), pages 312–320, Atlanta, US

"Introduction to Neural Networks: Anatomy | Machine Learning Crash Course | Google Developers." *Google*, Google, 10 Apr. 2018, developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/anatomy .