Sentiment analysis for Amazon.com reviews

Big Data in Media Technology (DM2583)
KTH Royal Institute of Technology, Stockholm

Levent Güner
leventg@kth.se

Emilie Coyne
ecoyne@kth.se

Jim Smit
jmsmit@kth.se

2019-03-01

Abstract

Sentiment analysis is a classification process whereby machine learning techniques are applied on text-driven datasets in order to analyze its sentiment, e.g. a message being positive or negative about a certain topic. We want to investigate if these sentiment analysis techniques are also feasible for application on product reviews form Amazon.com. In this study, different machine learning algorithms are compared, trained and tested on a dataset ($N = 60,000$) containing product reviews from Amazon.com which are randomly selected from dataset available from Kaggle containing 4 million reviews. The performance of three different algorithms were compared: Multinomial Naive Bayes (MNB), Linear Support Vector Machine (LSVM) and Long short-term memory network (LSTM). The LSTM resulted in the highest performance ($Accuracy = 0.90, AUC = 0.96$). Thereafter, to evaluate the LSTM model, it was applied on the remaining 3.94 million reviews from the Kaggle dataset, as well as on a new scraped dataset from Amazon.com containing reviews on products from different categories. This resulted in a very accurate classification, with the best results for reviews on furniture products ($Accuracy = 0.92$). In conclusion, LSTM networks are very suitable for classification of the sentiment on product reviews and the results do not change significantly for different categories. Further study is needed to investigate if the classification remains accurate when including more than two classes (e.g. introducing a neutral class).

Introduction

Online shopping has been growing for 20 years and many e-commerce websites such as Amazon, have been created to meet the increasing demand. Consequently, a specific product can be bought on several websites and the prices may vary. As customers usually want the best quality for the lowest price but can't directly check it, reviews from other customers seem to be the most reliable way to decide whether to buy the product or not. Therefore, sentiment analysis has proven essential to understand a product's popularity among the buyers all over the world.

Related works

Multiple studies about sentiment analysis on Amazon.com reviews have been done [1]. These studies used conventional Machine Learning (ML) like Naive Bayesian (NB), Support Vector Machine (SVM), decision trees or logistic regression, which resulted in relatively good performances ($accuracy > 0.90$).

- A sentiment analysis of reviews of Amazon beauty products has been conducted in 2018 by a student from KTH [2] and he got accuracies that could reach more than 90% with the SVM and NB classifiers. He found that SVM was performing better than NB for a large amount of data. He also focused on summaries of the reviews which are more informative and got higher accuracy than with the complete reviews.

- Xing Fang and Justin Zahn analyzed different categories of Amazon products (beauty, book, electronic, and home) [3] with 3 different classifiers: NB, SVM and Random Forest. They reached the conclusion that Random Forest usually provided them with more accurate results. They also found that SVM was performing better than NB for larger data sets.

- Some work has also been done about binary classification with LSTM network.

  – Zhenxiang Zhou and Lan Xu analyzed the usefulness of Amazon food reviews [4] with LSTM and feed-forward neural (FFN) networks. The results have shown that LSTM outperformed FFN, and that the accuracy was quite good ($\simeq 80\%$).

  – Reviews of Amazon books have also been analyzed, using LSTM algorithm in 2017[5] . They compared two recurrent neural networks (RNN): Gated Recurrent Unit (GRU) and Bidirectional LSTM. The bag-of-word algorithm was used for feature extraction. With a data set of more than 210 000 reviews, they got the best accuracy with the LSTM algorithm (86%).

The aim of this project is to investigate if sentimental analysis is feasible for the classification of product reviews from Amazon.com. Therefore, we will compare the performance of different classification algorithms on the binary classification (positive vs. negative) of product reviews from Amazon.com. Thereby, we want to investigate whether the category of products the reviews come from influence the performance of this classification. Once found the best performing classifier,it will be applied on new Amazon.com datasets containing reviews of different product categories and these results will be compared.

Method

Data acquisition

The dataset used for training consisted a big dataset (4 million reviews) available on Kaggle [6]. This Kaggle dataset consists of Amazon customer reviews (input text) and binary output labels (positive and negative), which were based on the star rating of the review. Here, a rating of 1 or 2 stars is labeled as negative while 4 or 5 were labeled as positive and 3 stars rated reviews (representing the neutral class) were excluded. The positive and negative class are both equally represented.
Furthermore, different pages of Amazon.com were searched for reviews and these were scraped from the website using a scraping script available in Kaggle. The following search terms were used: `['Laptop', 'Camera','Playmobil', 'Lego','Chair','Bed']`. This resulted in a dataset containing 3 different categories of Amazon products: Electronic devices, Toys and furniture. Table 1 shows furthers specifications of the all the used datasets.

Dataset	**Number of reviews**	**fraction of positive label**
Training set	40,000	0.50
Testing set	20,000	0.49
Evaluation sets		
Kaggle	3,940,000	0.50
Electronic devices	64,237	0.85
Toys	21,613	0.97
Furniture	146,653	0.90

Table 1: Data specifications

Data pre-processing

The raw review data is cleaned for different elements, which could worsen the performance of the classifier. This is done by:

- Removal of HTML tags.
- Filtering every symbol except for letters (a-z) and numbers (0-9).
- Filtering out every word with length of 3 symbols or lower.

Feature extraction
After data cleaning, feature extraction methods are applied to convert the text data into numerical data.
For the classification with MNB and LSVM, a TF-IDF vectorizer is used. Whereas for the classification with LSTM, a method called tokenization is applied.
The tokenization process consists of three parts which are fitting on training set, converting text to sequences and padding sequences.
The fitting on the training set is based on a 'maximum features' parameter, which represents the maximum number of unique words to be recognized in the total of reviews. Converting text to sequences is separating the words one by one and converting them to

unique integers.
An example for a review: "It is a really really good product" will be tokenized as `['it','is','a','really','really','good','product']` and then converted to `[1,2,3,4,4,5,6]`.
So `1` represents `it`, `4` represents `really`...
In a new sentence, these words will be replaced according to this mapping.
Padding sequences is the process of making every sequenced array the same length, which means basically adding zeros to the arrays.
The reason that TF-IDF vectorization is applied for MNB/SVM is that certain data is lost. Whereas with tokenization, the numerical data represents the whole sentence, which is more convenient for a recurrent neural network like LSTM.

Classifiers

To investigate which machine learning modality performs best on the classification of Amazon.com reviews, three different machine learning modalities were trained.

- Multinomial naive Bayesian (MNB)
- Linear support vector machine (LSVM)
- Long short-term Memory (LSTM) networks.

Since Naive Bayes and SVM are considered as conventional algorithms and are widely used in the field of sentimental analysis, we will use these classifiers as benchmarks.
LSTM is a newer technique and is shown to have a high potential for a good performance in sentiment analysis. For the LSTM networks, We created our model with Keras library, which consists of 4 layers:

- Embedding: Reducing the size of inputs
- Spatial dropout: To prevent from overfitting
- LSTM: Long Short Term Memory layer, which is the RNN
- Dense: To convert LSTM outputs to binaries

Training / Testing process

To keep the computational expense limited, a randomly selected subset out of the training dataset, consisting 60.000 reviews, was extracted. From this 60.000 reviews, 40.000 were used to train the classifiers whereas the remaining 20.000 were used to test their performances.

Evaluating metrics

To decide which classification algorithm performed the most accurate on the test set, we used several performance metrics:

- Accuracy: it compares the predicted general sentiment (positive or negative) to the real one, which was determined based on the stars.

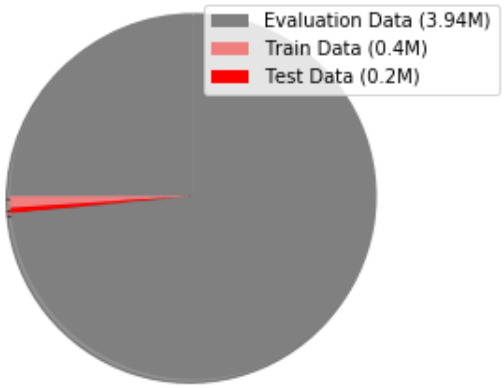


Figure 1: Kaggle dataset distribution

- AUC: The Area Under Curve (AUC) is a metric where the False Positive Rate (FPR) and True Positive Rate (TPR) are combined into one single metric. First, the FPR ad TPR are computed with many different thresholds for the classification algorithm. These FPRs and TPRs are parametrically plotted in a single graph, which results in the Receiver Operating Characteristic (ROC) curve. Finally, the metric we consider is the Area of this curve, which we call AUROC or AUC.
- Precision: this is the ration between True Positives and the sum of True Positives and False Positive reviews. It tells us how accurate we are about saying that a review is positive.
- Recall: this is the ration between True Positives and the sum of True Positives and False Negatives.
- F1-score: this is the harmonic mean of the precision and the recall.

Evaluation process

To evaluate the performance of the trained classifier, the classifier with the best performance on the test data is applied on new, unseen reviews. This evaluation set was collected by picking remaining 3.94 million reviews of the Kaggle dataset [6](See Figure 1).
We applied the best performing classifier after 3.94 million reviews, also on the self-scraped dataset from Amazon.com. This is done for the three different product categories separately.

Results

Training results

In Table 2, the results on the test data are represented for the different machine learning algorithms.

ML algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1-score**
Linear SVM	0.86	0.93	0.86	0.86	0.86
Multinomial NB	0.85	0.93	0.87	0.82	0.84
LSTM network	0.90	0.96	0.92	0.87	0.90

Table 2: Results for test data

Dataset	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1-score**
3.94M Review Set	0.90	0.96	0.75	0.87	0.90
Toys	0.91	0.95	0.99	0.91	0.95
Tech Products	0.90	0.95	0.98	0.91	0.94
Furnitures	0.92	0.95	0.98	0.93	0.95

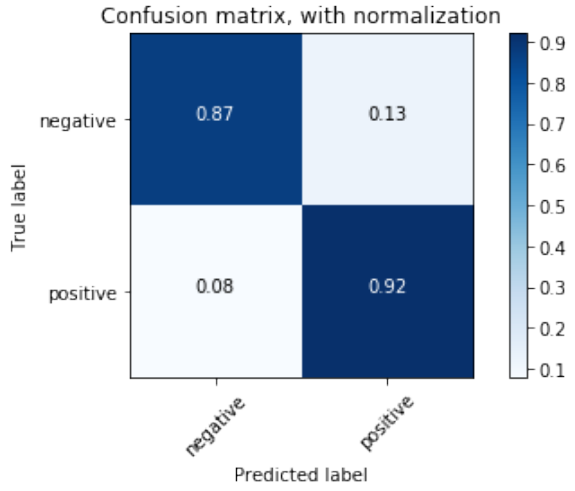
Table 3: Results for evaluation data

As showed in Table 2, we get the highest accuracy with the LSTM network. The other performance metrics, such as AUC or F1-score, are also higher with the LSTM network compared with the other algorithms. Therefore, we consider this alogorithm as the most suitable for the sentiment analysis on Amazon reviews and used it to classify the reviews of the evaluation datasets. These results are given in Table 3 and the confusion matrices are shown in Figure 2.

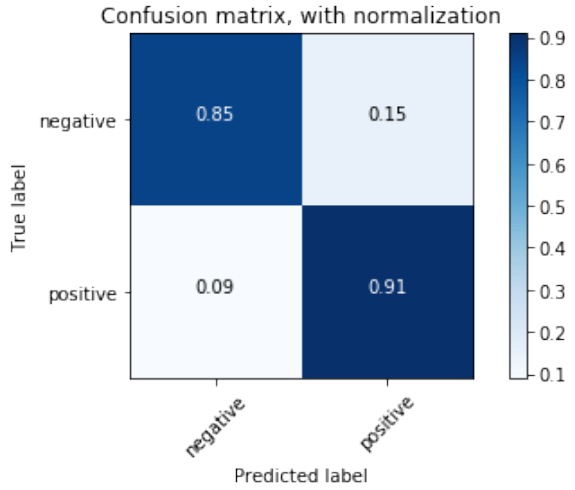
Discussion

To summarize our work, we selected 60000 random reviews between 4 million reviews, cleaned and tokenized them, created different models and selected the best. After this process, we evaluated them with the remaining 3.94 million reviews and scraped $\sim$ 230000 real time reviews from Amazon.com to answer our question "Are sentiment analysis methods feasible for Amazon.com reviews?".

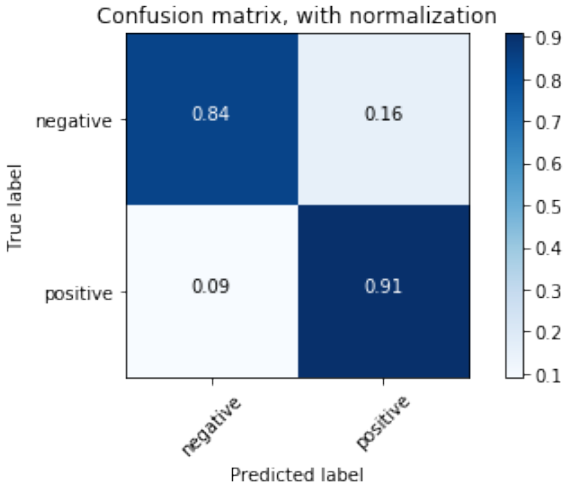
Test and evaluation results
The results from Multinomial Naive Bayes (MNB) and Linear Support Vector Machine (LSVM) were also satisfying, however, since TF-IDF vectorization limits the data and tokenization is not efficient for those classifiers, tokenizing the sentences and training with neural networks gave the best results. Here we can see that not only the classification method, but also feature extraction has an important role in the process. Different types of Neural Networks may give more accurate results; however, when we see previous works and researches, we can say that Long-Short Term Recurrent Neural Networks are working pretty much for sentiment classification. Since the results are very comparable



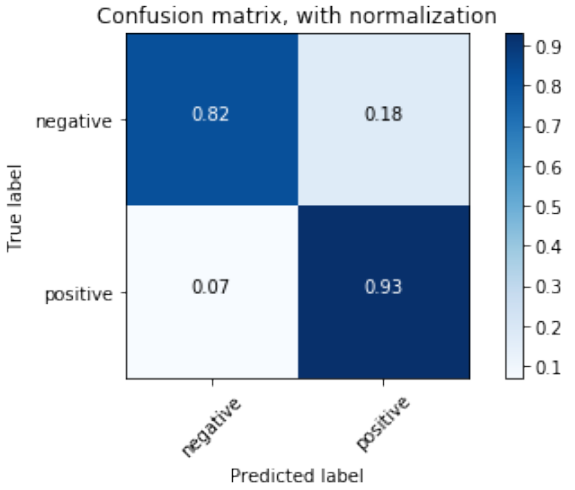
(a) Kaggle Evaluation Set



(b) Scraped Amazon Reviews - Toys



(c) Scraped Amazon Reviews - Tech Devices



(d) Scraped Amazon Reviews - Furniture

Figure 2: Confusion Matrices for LSTM classifications on the evaluation data sets

(all ≥ 0.90 accuracy) on different datasets, containing randomly selected reviews (Kaggle dataset) and reviews from different product categories collected by searching the Amazon.com webpage for different searching terms, it is very unlikely that the performance of the LSTM depends strongly on the category of the product reviews.

Study limitations

In this study, the assumption is made that the amount of stars corresponds with the sentiment in the review. However, it could be the case that a very positive review is given one star or vice versa, polluting the dataset. This could have been checked by using another (labeled) dataset (e.g. the VADER dataset) and classify the the reviews by a model trained on this external data. In this way you could check whether the classification corresponds with the labels (or stars) which are given to the reviews.

Another important process in machine learning, is cross-validation. This has the aim to eliminate the risk that an 'easy-to-classify' subset of your data is selected as the test data by chance, biasing your results. For the training and testing phase in this project, this has not been implemented.

Furthermore, the different classifiers have certain hyperparameters which can be tuned to optimize their performance on the specific data you are dealing with (e.g. 'C' and 'gamma' parameters for the SVM). Several approaches, like a gridsearch, can be used to tune these parameters. In this study, parameters for SVM and MNB are set to their defaults and their results may have been even better when optimized.

Future studies
For future studies it may be interesting to check how well the LSTM algorithm will perform when more classes are added to the classification. In this study, the reviews with 3 stars were excluded from the classification, which could have been added as well as a 'neutral' class. To extend this idea, reviews can even be divided into five different classes according to the amount of stars which are given. In this case, it will not be easy to find or collect a suitable dataset, since the reviews on Amazon.com are mostly positive, with the vast majority ranked with five stars. However, when enough data is collected for training, extending the work for predicting more classes, with a correct model and enough data is expected to give more accurate results. Different combinations of feature extraction and creating models for those methods may give better results.

Conclusion

As the results on the test data shows, LSTM networks are the most suitable for binary sentiment analysis on Amazon.com product reviews. Based on the results on the evaluation datasets, we can conclude that LSTM performs very well ($accuracy > 0.90$) for binary classification, and that does not depend strongly on the type of product where the reviews come from. As it can be seen clearly from confusion matrices in Figure 2, the LSTM network both performs accurate results for positive and negative classes. Since the training dataset is also balanced, getting balanced results from both classes shows the model's reliability.
In conclusion, LSTM networks are very suitable for classification of the sentiment on product reviews. These results do not depend on the type of product where the reviews are given to. Since we get $\sim 90\%$ accuracy with just training 1% of data, we can say that sentimental analysis methods are feasible with Amazon review data.

References

[1] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," *2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018*, no. May, pp. 1–6, 2018.

[2] S. Paknejad, "Sentiment classification on Amazon reviews using machine learning approaches," 2018.

[3] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, no. 1, 2015. [Online]. Available: http://dx.doi.org/10.1186/s40537-015-0015-2

[4] Z. Zhou and L. Xu, "Amazon Food Review Classification using Deep Learning and Recommender System," *Stanford University*, pp. 1–7, 2009. [Online]. Available: https://cs224d.stanford.edu/reports/ZhouXu.pdf

[5] J. Nowak, A. Taspinar, and R. Scherer, "LSTM recurrent neural networks for short text and sentiment classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10246 LNAI, pp. 553–562, 2017.

[6] "Amazon Reviews for Sentiment Analysis." [Online]. Available: https://www.kaggle.com/bittlingmayer/amazonreviews#train.ft.txt.bz2