Sentiment Classification of Amazon Reviews

Yina Bao

DATS6501 Data Science Capstone Project

Dec 7, 2018

Motivation

- In recent years, electronic commerce more and more dominates the market, and there has been a huge increase of interest from brands, companies and data scientists in sentiment analysis and the application to find the business intelligence insight.

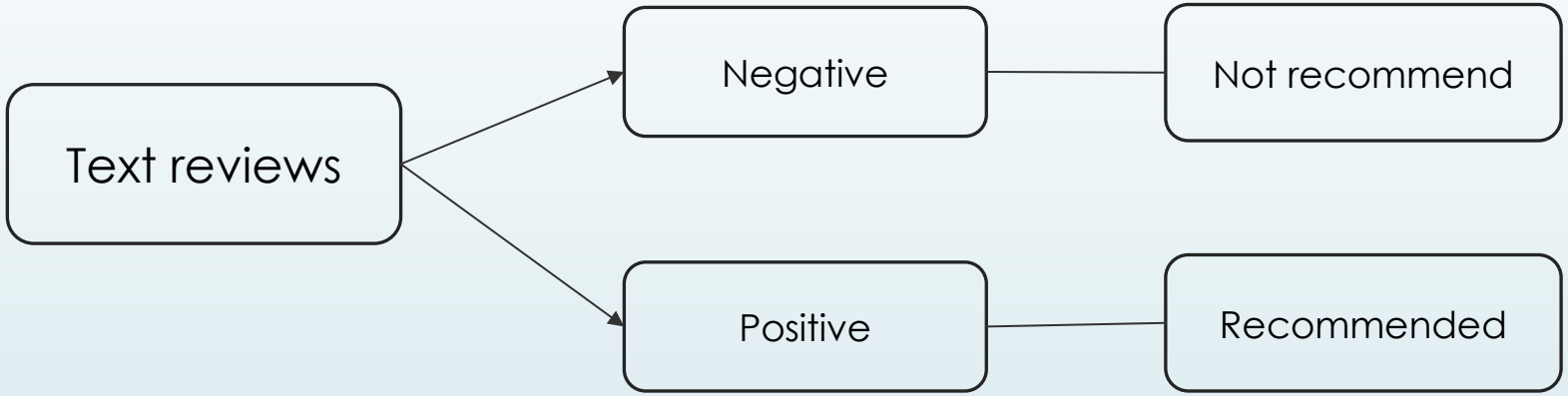
- Recent study from Zendesk mentioned that 45% of people share negative customer service experience and 30% share positive customer service experience via social media, which shows a high demand for mining the information and extracting the opinion and meaning for further analysis.

- The model of classifying Amazon reviews can transfer to some text without labels, in order to give a sentiment suggestion (such as whether recommend a video on YouTube based one the comments).

Objective

- Understanding customer's sentiment based on descriptive text reviews.

  Review classification:

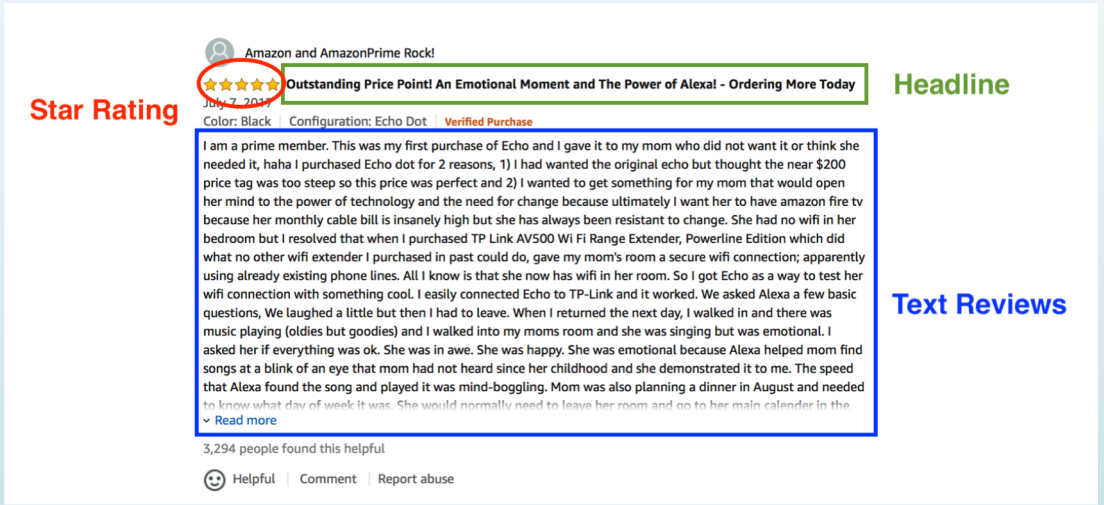


Dataset

- Amazon product reviews dataset consists of star ratings (1-5), the headline and the descriptive customer reviews.
- The original dataset is pre-split into a testing set and training set. Training set contains 3,000,000 reviews and testing set contains 650,000 reviews.
- Link to Kaggle data page: https://www.kaggle.com/bittlingmayer/amazonreviews

  Link to dataset: https://drive.google.com/drive/folders/0Bz8a_Dbh9Qhbfll6bVpmNUtUcFdjYmF2SEpmZUZUcVNiMUw1TWN6RDV3a0JHT3kxLVhVR2M

1	mens ultrasheer	This model may be ok for sedentary types
4	Surprisingly delightful	This is a fast read filled with unexpected
2	Works, but not as advertised	I bought one of these chargers..the instru
2	Oh dear	I was excited to find a book ostensibly ab
2	ncorrect disc!	I am a big JVC fan, but I do not like this m
2	ncorrect Disc	I love the style of this, but after a couple
2	DVD menu select problems	I cannot scroll through a DVD menu that i
3	My 2 y/o grandson loves it!!	This movie with all of its animals really k
5	A Cookbook Every Baker Should Own	I found a copy of this cookbook at a local
3	good basic	The book is a basic "how to" book for usin
3	nice screen for a nice price but....	I compared a few different flat panels wi
3	Poor maps, no hostels	It's a good book, but the maps are not ver
2	Profound then. Truly horrible now.	The narrative style of this work by famous
1	A complete Bust	This game requires quicktime 5.0 to work
5	Barbie as Rapunzel: A Creative Adventu	I purchased this software for my 5 year ol



Preprocessing – clean data

- Random select 200,000 data from original training set and 50,000 data from original testing set to create a smaller balanced dataset.
- Cleaning the data:
  - Covert each word to lower case
  - Remove web link
  - Remove @ social media account
  - Remove non-alphabetic characters
- Export as a csv file for saving the cleaning time.

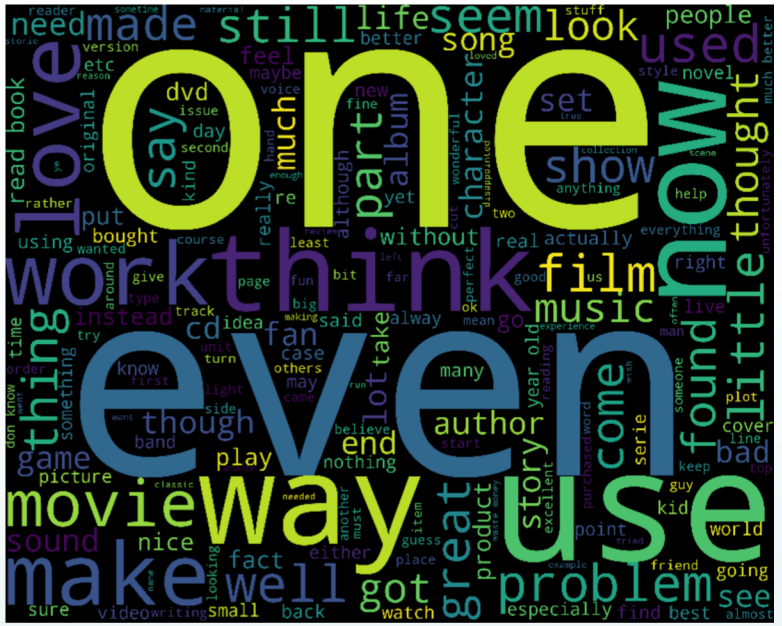
```
i = re.sub(r'http\S+', '', i)
i = re.sub(r'@\S+', '', i)
i = re.sub('[^a-zA-Z]', ' ', i)
```

```
The Ultimate Movie Review! - [...] - @tss5078

What I need is something like this: http://www.amazon.com/General-UV513AB-Digital-Ultraviolet-Measurement/dp/B002JOR0JO/ref=rsl_mainw_dpl?ie=UTF8&m;=ATVPDKIKX0DERBut it's a lot more expensive...

This machine is good to buy as a starting point but as is the fence is way too small to support reasonable sized work pieces. So, buy it, bolt a steel plate about 6 x 12 to the fence, about 1/4" thick, and enjoy. Works fine otherwise.
```

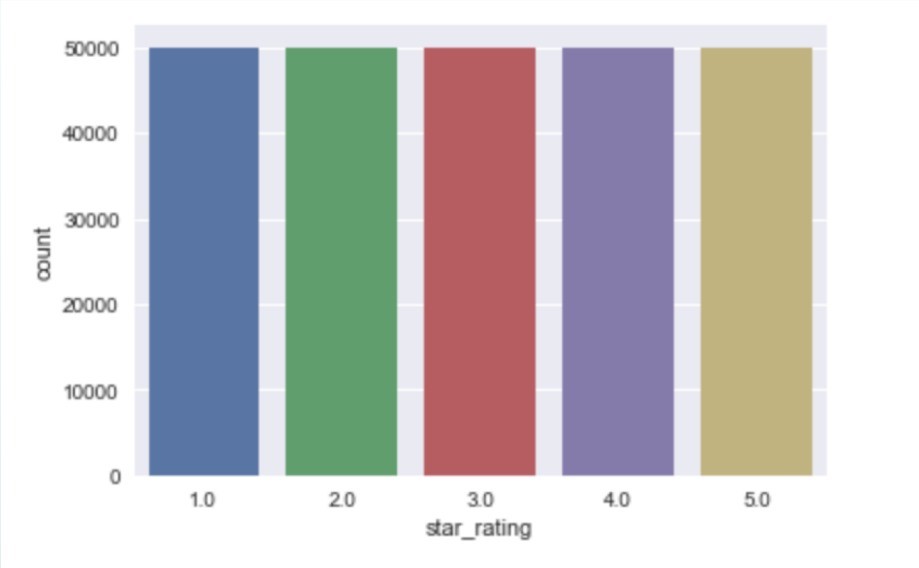
Examples after cleaning

```
the ultimate movie review

what i need is something like this it s a lot more expensive

this machine is good to buy as a starting point but as is the fence is way too small to support reasonable sized work pieces so buy it bolt a steel plate about x to the fence about thick and enjoy works fine otherwise
```

Word Cloud for headline



Word Cloud for text_reviews



Balanced data!

Preprocessing

- Inputs:
  - Removing stop-words by *nltk.corpus.stopwords* , and stemming by *nltk.stem.SnowballStemmer*
  - Bag of words – numerical representation of text
    Using *TfidfVectorizer* (term frequency–inverse document frequency) and *fit_transform* method in *sklearn* to fit the model of bag of words to create the feature vectors.
- Labels:
  Positive-1: contains star 4 and star 5
  Negative-0: contains star 1 and star 2
- Split 80% for training set and 20% for testing set.

Inputs comparison – Headline vs Reviews

F1-score	SVM	Naïve Bayes	Random Forest
Headline	0.7872 (+/- 0.0051)	0.7893 (+/- 0.0062)	0.7733
Reviews	0.8645 (+/- 0.0041)	0.8406 (+/- 0.0037)	0.8340

Classification by reviews has better performance.

Although headline contains lots of key words of sentiment, the headline is much shorter than reviews. Removing stop-words makes several empty inputs.

Use text_reviews column for following model comparison.

Models comparison

	MLP	CNN	SVM	Naïve Bayes	Random Forest
Accuracy	0.79	0.83	0.86	0.84	0.83
F1 score	0.79	0.83	0.86	0.84	0.83

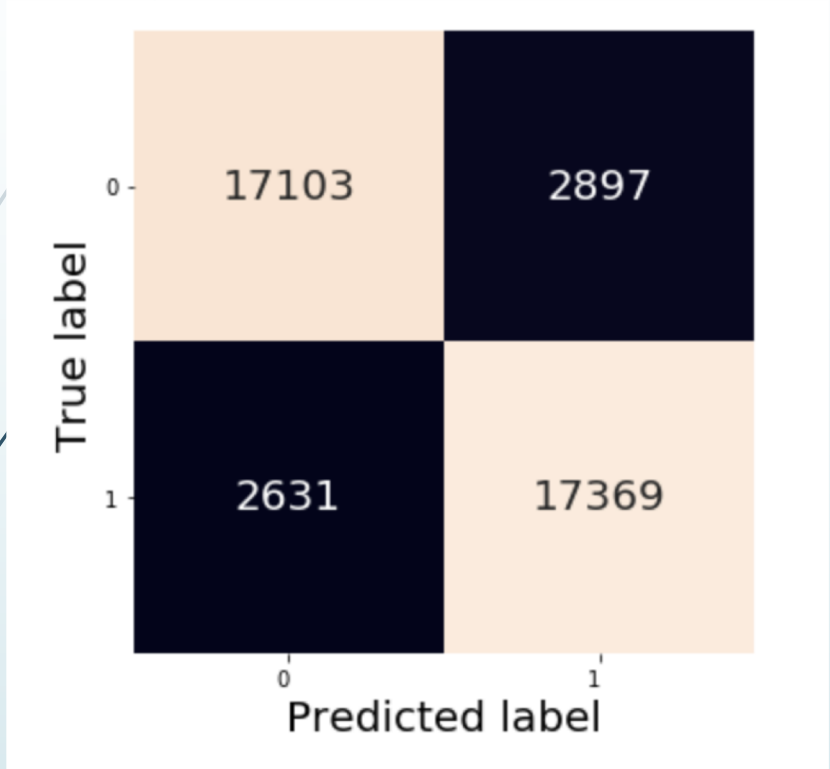
Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 915, 128)	2560000
flatten_1 (Flatten)	(None, 117120)	0
dense_1 (Dense)	(None, 256)	29982976
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 512)	131584
dropout_2 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 2)	1026
activation_1 (Activation)	(None, 2)	0

Total params: 32,675,586
Trainable params: 32,675,586
Non-trainable params: 0

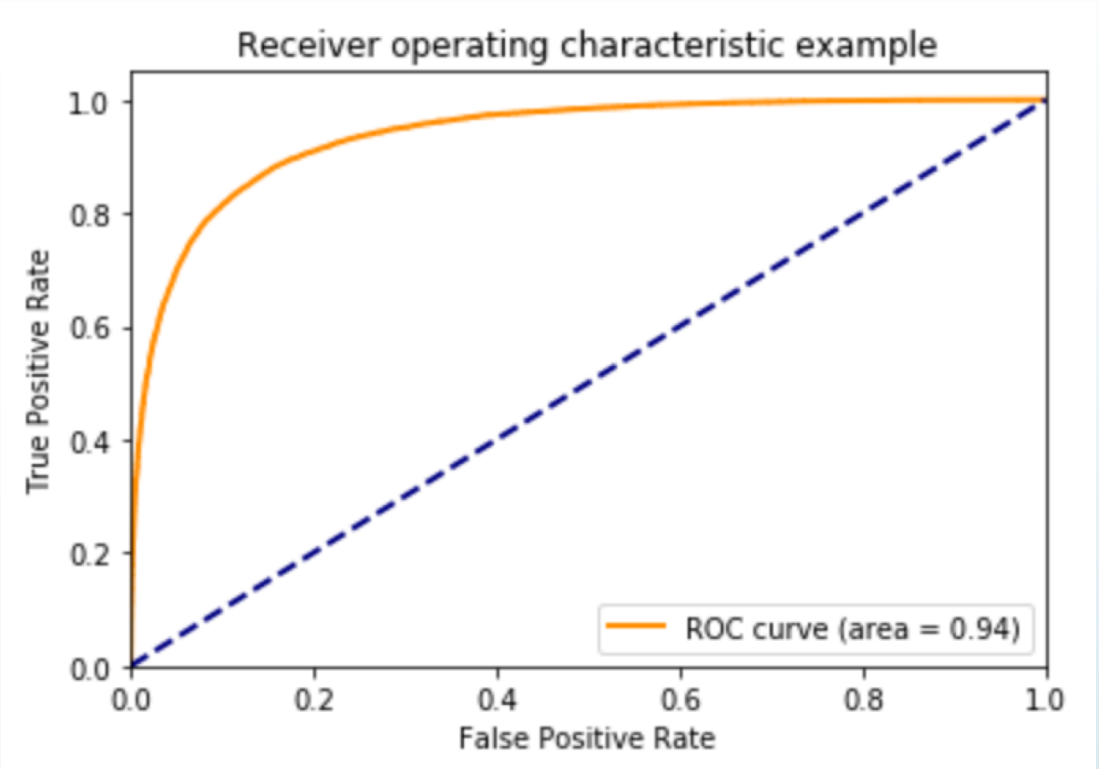
Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 915, 128)	2560000
conv1d_1 (Conv1D)	(None, 913, 256)	98560
global_max_pooling1d_1 (Glob	(None, 256)	0
dense_1 (Dense)	(None, 512)	131584
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 512)	262656
dropout_2 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 2)	1026
activation_1 (Activation)	(None, 2)	0

Total params: 3,053,826
Trainable params: 3,053,826
Non-trainable params: 0

Confusion Matrix by SVM model



ROC curve by SVM



See what's going on 5 classes…

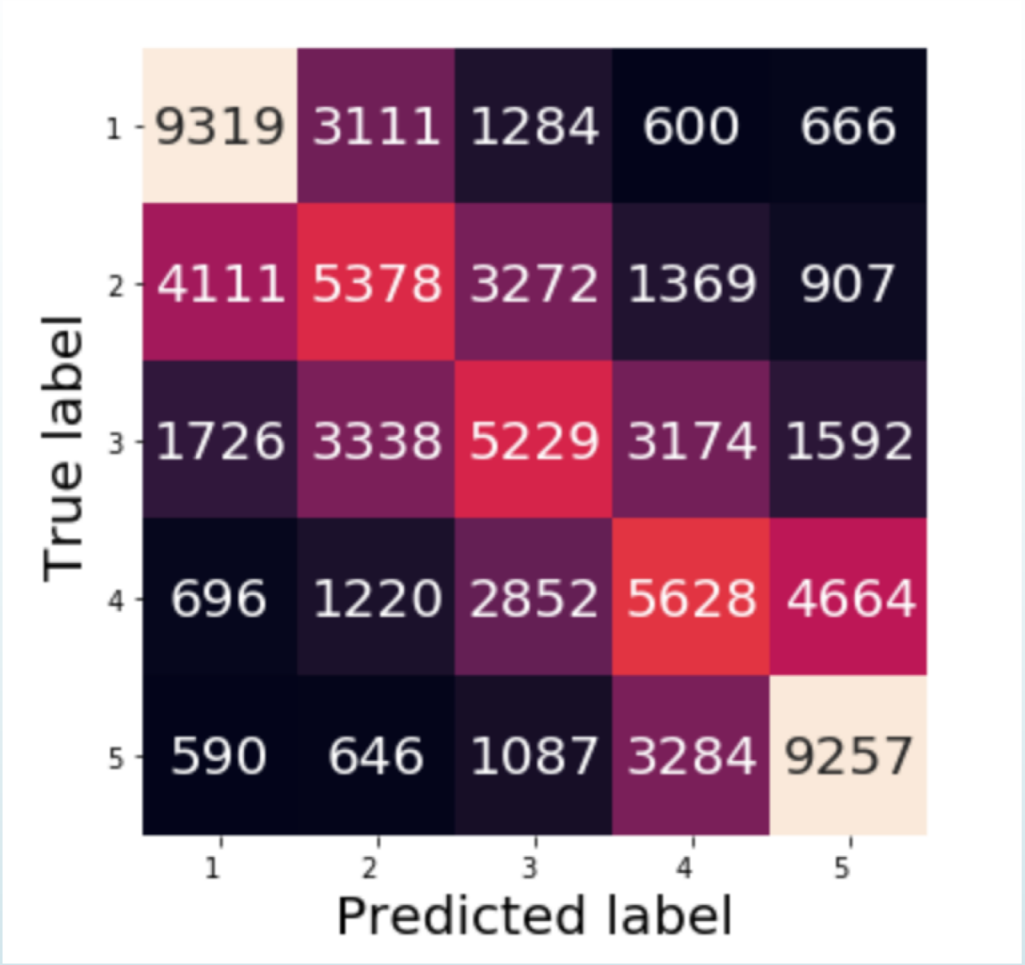
SVM

Accuracy: 0.46

F1-score: 0.45

Based on the number and color on confusion matrix, star 1 and star 5 have the highest true positive rate.

Include 3 star in either positive or negative will decrease the accuracy to 0.77



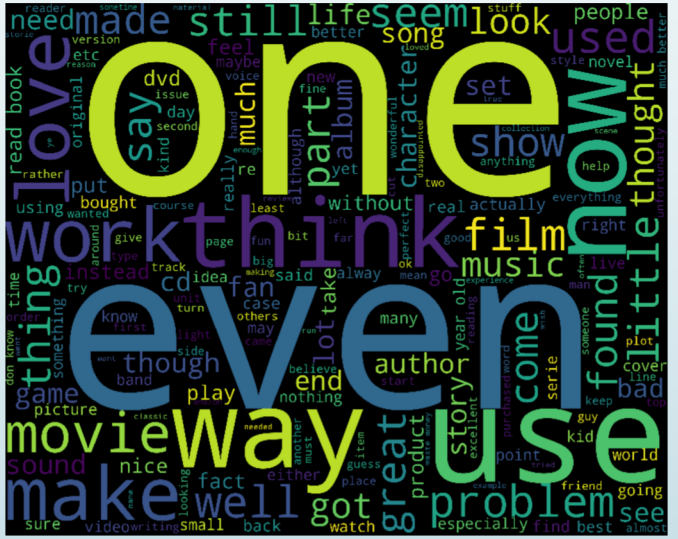
Confusion matrix of SVM in 5 classes

Summary

- SVM model has better results in sentiment classification
  - Using GridsearchCV to select the best hyper-parameter
- Build up more layers of MLP & CNN might improve the performance
- Problems of preprocessing:
  - Stopwords contain "sentiment" words, such as 'not'
  - Misspelling
  - Different languages
  - Bag of words not consider the order
- Natural language is "versatile", "elusory"…(such as sarcastic)

Future work

- Pre- trained word embedding:
  Glove, FastText as transfer learning
- Combine headline and text reviews
- Lemmatizer
- Classifying neutral reviews by our sentiment classification model, and clustering the neutral reviews to find some patterns (whether the positive-neutral reviews and negative-neutral reviews are clustered separately).
- Classification based on 5 classes



Thank you