Presentation at Radical.io, June 7, 2019

Introduction to Natural Language Processing (NLP): Sentiment Analysis on 515K Hotel Reviews

Xinru Cheng
GitHub: floraxinru

Introduction

Data Science

Goal: ***generate insights from data, to take data-driven actions***

The Data Science Process (iterative, team effort):

Acquire access and retrieve data

Prepare exploratory data analysis
pre-processing:
clean, integrate, package

Analyze choose techniques, build model

Report communicate insights

Act apply results

Introduction

Data Science

Goal: ***generate insights from data, to take data-driven actions***

The Data Science Process (iterative, team effort):

Acquire access and retrieve data

Prepare exploratory data analysis
pre-processing:
clean, integrate, package

Analyze choose techniques, build model

Report communicate insights

- -

Act apply results

Introduction

Data Science

Goal: ***generate insights from data, to take data-driven actions***

Machine Learning

The Data Science Process (iterative, team effort):

Acquire access and retrieve data

Prepare exploratory data analysis
pre-processing:
clean, integrate, package

Analyze choose techniques, build model

Report communicate insights

Act apply results

Introduction

Data Science

Goal: ***generate insights from data, to take data-driven actions***

Machine Learning

Supervised (target available)	**Unsupervised** (target unavailable)
Classification	Cluster Analysis
Regression	Association Analysis

The Data Science Process (iterative, team effort):

Acquire access and retrieve data

Prepare exploratory data analysis
pre-processing:
clean, integrate, package

Analyze choose techniques, build model

Report communicate insights

Act apply results

Natural Language Processing (NLP)

- Use algorithms and data techniques to analyze, understand and derive meaning from human language effectively and efficiently
- Difficult computationally because human language is ambiguous — need context and the ability to link concepts
- Applications:

  - speech recognition engines, automatic translators, chatbots, generate keywords and trending topics, preventing spam, preventing fake news, identifying product mentions on Twitter, diagnosing medical reports and classifying legal texts
- Sentiment analysis: identification of attitude or emotion in a text

Motivation

During the last decade, we have relied increasingly heavily on online ratings and reviews when making decisions, especially when travelling to a new destination.

In this project, I am interested in looking for words that are strong indicators of positive or negative hotel reviews through natural language processing and sentiment analysis.

This could provide valuable insight to hotel management as well as similar websites collecting ratings to improve their performance and better target certain customers. It might also help fellow travellers understand which words are the most effective when leaving a review for their next stay.

Dataset

Acquire

Prepare

Analyze

Report

Act

My dataset is the "515K Hotel Reviews Data in Europe" dataset on Kaggle (https:// www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe).

The dataset is a .CSV file of size 48MB, containing mostly text.

The positive and negative reviews are already in columns. The reviews are all in English, collected from Booking.com from 2015 to 2017.

The dataset contains 515738 reviews for 1493 luxury hotels in Europe.

Dataset

First two rows (first two reviewers):

	Hotel_Address	Additional_Number_of_Scoring	Review_Date	Average_Score	Hotel_Name	Reviewer_Nationality	Negative_Review	Review_Total_Negative_Word_Counts	Total_Number_of_Reviews	Positive_Review	Review_Total_Positive_Word_Counts
0	s Gravesandestraat 55 Oost 1092 AA	194	8/3/2017	7.7	Hotel Arena	Russia	I am so angry that i made this post available...	397	1403	Only the park outside of the hotel	11
1	s Gravesandestraat 55 Oost 1092 AA	194	8/3/2017	7.7	Hotel Arena	Ireland	No Negative	0	1403	No real complaints the hotel was great	105

17 Columns total

- Total_Number_of_Reviews_Reviewer_Has_Given
- Reviewer_Score
- Tags
- days_since_review
- lat
- lng

Data Preparation and Cleaning

The dataset did not require much cleaning before analysis. All the reviews are in English. The data uploader stated that punctuation was already removed, and all reviews were converted to lowercase.

The data preparation and cleaning I performed include selecting the columns for positive and negative reviews and filter out **stopwords** before sentiment analysis.

stopwords: *words like "the", "that", "is" that don't help with identifying the context*

Research Questions

Questions I aim to answer using this dataset include:

1. Can we perform sentiment analysis on the positive and negative reviews, to find out which words have the largest effect on predicting the review outcome?
2. Do experienced travellers tend to leave reviews that are more negative and have lower ratings?
3. Is there a correlation between reviewer nationality and high ratings?

Methods

Bag-of-words model

- The simplest model for analyzing text—think about text as an unordered collection of words.

generally allows us to infer the topic or the sentiment of the text

can build features from this model to be used by a classifier

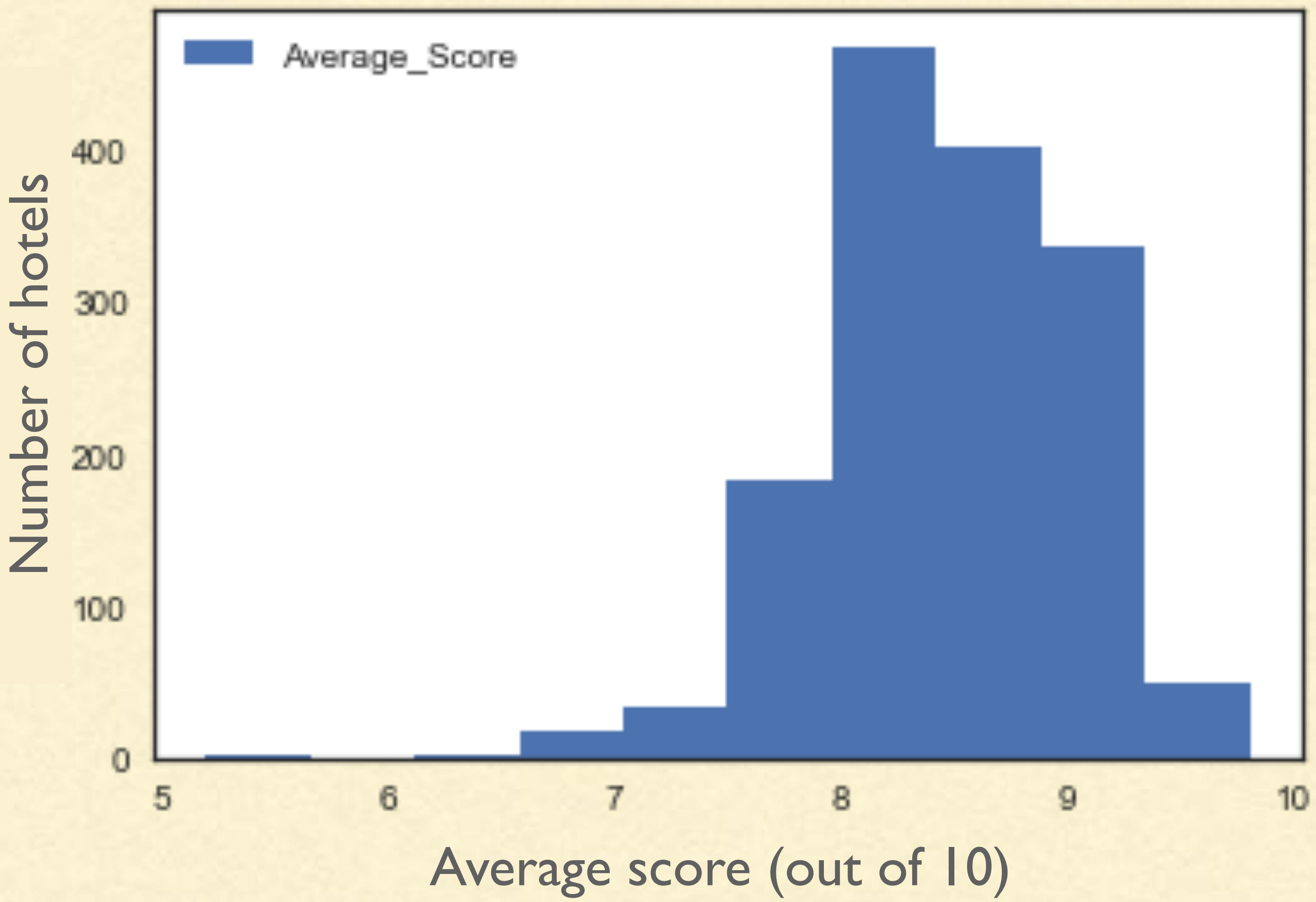
appropriate for finding the words most indicative of a positive or negative review, because for sentiment analysis we just need the *number of occurrences* of each word, not the *order* they are in

Naive Bayes Classifier

- One of the simplest supervised machine learning classifiers, it can be trained on 80% of the data to learn what words are generally associated with positive or with negative reviews

Findings

Average score of 1493 hotels



I first explored the dataset, and discovered that the average score of the 1493 European luxury hotels are mostly between 7.6 and 9.2.

The average scores range between 5.2 and 9.8, with a mean of 8.4.

Findings

I tokenized the text in the positive and negative reviews, and trained the classifier on 80 percent of the data, then tested it on the remaining 20 percent.

Regarding my main research question, my analysis has shown that we can use sentiment analysis of natural language processing (NLP) to make successful predictions on the positive and negative reviews on this dataset.

In particular, building a bag-of-words model to use with the Naive Bayes Classifier produced a training set accuracy of **93.5** percent, and a testing accuracy of **92.5** percent, both significantly larger than the estimated human accuracy of **80** percent. This indicates the Naive Bayes Classifier is a good method to use for this analysis.

Findings

I found the most informative features, which are the words that best identify a positive or a negative review, or the words that had the greatest effect on the prediction accuracy.

It is interesting to note that the most informative words for positive reviews tend to refer to the ***hotel staff (Friendly, Helpful, Efficient), room (Comfy, Spacious, Comfortable), and location (Convenient, Conveniently, Convenience)***, while the most informative words for negative reviews seem to refer mostly to problems with the ***facilities or amenities (unstable, Thin, Charged, Unusable, Lack, unreliable, damaged, Loud, Noisy, Smelly, Missing, loudly)***.

This could be valuable insight for the reviewed hotels that are looking for areas to improve, in order to increase their ratings and attract more customers.

Identifying the words that best indicate sentiment in the reviews

(number of reviews versus 1)

```
Most Informative Features
             Negative = 1                    neg : pos    =  22605.9 : 1.0
             Positive = 1                    pos : neg    =  11601.8 : 1.0
                Comfy = 1                    pos : neg    =    234.6 : 1.0
          Outstanding = 1                    pos : neg    =    211.7 : 1.0
             Friendly = 1                    pos : neg    =    208.5 : 1.0
             Spacious = 1                    pos : neg    =    184.5 : 1.0
            Brilliant = 1                    pos : neg    =    168.8 : 1.0
              History = 1                    pos : neg    =    154.3 : 1.0
             Charming = 1                    pos : neg    =    153.7 : 1.0
          Beautifully = 1                    pos : neg    =    133.4 : 1.0
           Convenient = 1                    pos : neg    =    132.4 : 1.0
              Helpful = 1                    pos : neg    =    125.3 : 1.0
            Excellent = 1                    pos : neg    =    121.8 : 1.0
            Fantastic = 1                    pos : neg    =    116.1 : 1.0
          Comfortable = 1                    pos : neg    =    114.6 : 1.0
            Delicious = 1                    pos : neg    =    109.0 : 1.0
            Luxurious = 1                    pos : neg    =    108.3 : 1.0
             unstable = 1                    neg : pos    =    108.3 : 1.0
                 Thin = 1                    neg : pos    =    103.7 : 1.0
         Conveniently = 1                    pos : neg    =    103.7 : 1.0
```

Identifying the words that best indicate sentiment in the reviews

(number of reviews versus 1)

```
Most Informative Features
            Negative = 1              neg : pos    =  22605.9 : 1.0
            Positive = 1              pos : neg    =  11601.8 : 1.0
               Comfy = 1              pos : neg    =    234.6 : 1.0
         Outstanding = 1              pos : neg    =    211.7 : 1.0
            Friendly = 1              pos : neg    =    208.5 : 1.0
            Spacious = 1              pos : neg    =    184.5 : 1.0
           Brilliant = 1              pos : neg    =    168.8 : 1.0
             History = 1              pos : neg    =    154.3 : 1.0
            Charming = 1              pos : neg    =    153.7 : 1.0
         Beautifully = 1              pos : neg    =    133.4 : 1.0
          Convenient = 1              pos : neg    =    132.4 : 1.0
             Helpful = 1              pos : neg    =    125.3 : 1.0
           Excellent = 1              pos : neg    =    121.8 : 1.0
           Fantastic = 1              pos : neg    =    116.1 : 1.0
         Comfortable = 1              pos : neg    =    114.6 : 1.0
           Delicious = 1              pos : neg    =    109.0 : 1.0
           Luxurious = 1              pos : neg    =    108.3 : 1.0
            unstable = 1              neg : pos    =    108.3 : 1.0
                Thin = 1              neg : pos    =    103.7 : 1.0
          Conveniently = 1            pos : neg    =    103.7 : 1.0
```

Identifying the words that best indicate sentiment in the reviews

(number of reviews versus 1)

```
Most Informative Features
              Negative = 1              neg : pos    =  22605.9 : 1.0
              Positive = 1              pos : neg    =  11601.8 : 1.0
                 Comfy = 1              pos : neg    =    234.6 : 1.0
           Outstanding = 1              pos : neg    =    211.7 : 1.0
              Friendly = 1              pos : neg    =    208.5 : 1.0
              Spacious = 1              pos : neg    =    184.5 : 1.0
             Brilliant = 1              pos : neg    =    168.8 : 1.0
               History = 1              pos : neg    =    154.3 : 1.0
              Charming = 1              pos : neg    =    153.7 : 1.0
           Beautifully = 1              pos : neg    =    133.4 : 1.0
            Convenient = 1              pos : neg    =    132.4 : 1.0
               Helpful = 1              pos : neg    =    125.3 : 1.0
             Excellent = 1              pos : neg    =    121.8 : 1.0
             Fantastic = 1              pos : neg    =    116.1 : 1.0
           Comfortable = 1              pos : neg    =    114.6 : 1.0
             Delicious = 1              pos : neg    =    109.0 : 1.0
             Luxurious = 1              pos : neg    =    108.3 : 1.0
              unstable = 1              neg : pos    =    108.3 : 1.0
                  Thin = 1              neg : pos    =    103.7 : 1.0
          Conveniently = 1              pos : neg    =    103.7 : 1.0
```

Conclusions

Can we perform sentiment analysis on the positive and negative reviews, to find out which words have the largest effect on predicting the review outcome?

—Yes, the Naive Bayes Classifier produced a testing set prediction accuracy of **92.5** percent

The most informative words indicating a review to be positive or negative were found for this dataset. **Positive** reviews reflect more on **hotel staff and location**, while highly **negative** reviews tend to focus on **facilities**.

I did not find a correlation between experienced travellers and their review scores, or reviewer nationality and scores. It is possible the relationships exist for a larger dataset, or different types of hotels.

Limitations

The algorithm itself requires 80% of training data, so might not work well for a smaller dataset.

Certain inherent limitations of this dataset include the fact that it only contains English reviews collected from one website (Booking.com), and that the hotels are limited to luxury hotels in Europe.

For future work it might also be worthwhile to implement a spell-check mechanism for typos that appear in the reviews.

Further Work

This project is based on my Final Project for the Python for Data Science course on edx, first submitted in Dec. 2017. Further work on this dataset which would provide more insights include:

- building a **regression** model to predict ratings based on certain words
- filtering out reviews that could be misleading (“no negatives”) to increase prediction accuracy
- improving and exploring visualizations (such as using a Folium map to visualize the geographic location of hotels and nationalities of reviewers)

A revolutionary approach for NLP was developed in 2018, using neural networks and inductive transfer learning for text classification[1]. It would be very interesting to apply it to this dataset.

[1] https://arxiv.org/abs/1801.06146

Other Projects (in-progress)

- Machine Learning and Deep Learning Project - Unsupervised Learning with Unstructured Data
    - Building a Recipe Recommendation System (Group Project)
- Data Cleaning and Data Visualization Projects
    - Visualizing World Abortion Laws and Regulations
    - Multiple Topics (Mental Health in Tech Surveys, Global Waste Management, Employment Outcomes for Women in STEM; Topics in Computational Social Science and AI Ethics)

LinkedIn: Xinru Cheng, GitHub: floraxinru

References

DSE200X Python for Data Science MOOC Jupyter Notebooks (UCSanDiego on edx, 2017)

Python packages - Matplotlib, Pandas, SciPy and Seaborn documentations

StackOverflow for debugging code

Kaggle kernels: (https://www.kaggle.com/benhamner/python-data-visualizations, https://www.kaggle.com/alexisclt/who-s-improving-who-s-doing-worse/notebook, https://www.kaggle.com/janpreets/where-to-stay-in-europe)

USMFIT paper: Jeremy Howard and Sebastian Ruder, https://arxiv.org/abs/1801.06146 (2018)

New NLP Approach: ULMFiT

A revolutionary approach for NLP was developed in 2018, called **Universal Language Model Fine-tuning for Text Classification (ULMFiT)**

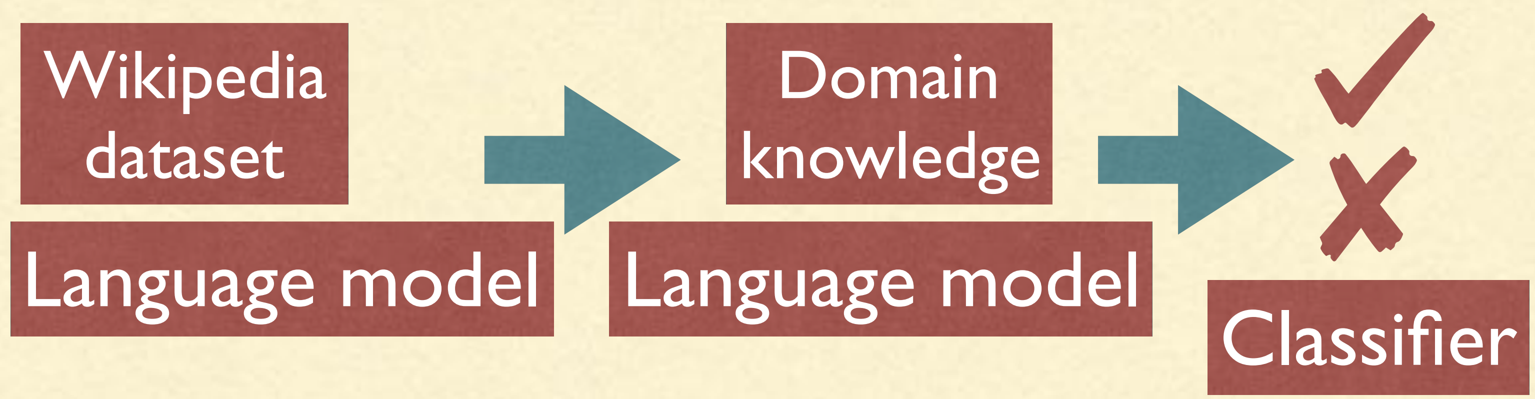
Using ***neural networks*** and ***transfer learning*** to train a language model, then use it as a classifier. A "***language model***" is any model that learns to predict the next word of a sentence

https://arxiv.org/abs/1801.06146

New NLP Approach: ULMFiT

A revolutionary approach for NLP was developed in 2018, called **Universal Language Model Fine-tuning for Text Classification (ULMFiT)**

Using ***neural networks*** and ***transfer learning*** to train a language model, then use it as a classifier. A "***language model***" is any model that learns to predict the next word of a sentence

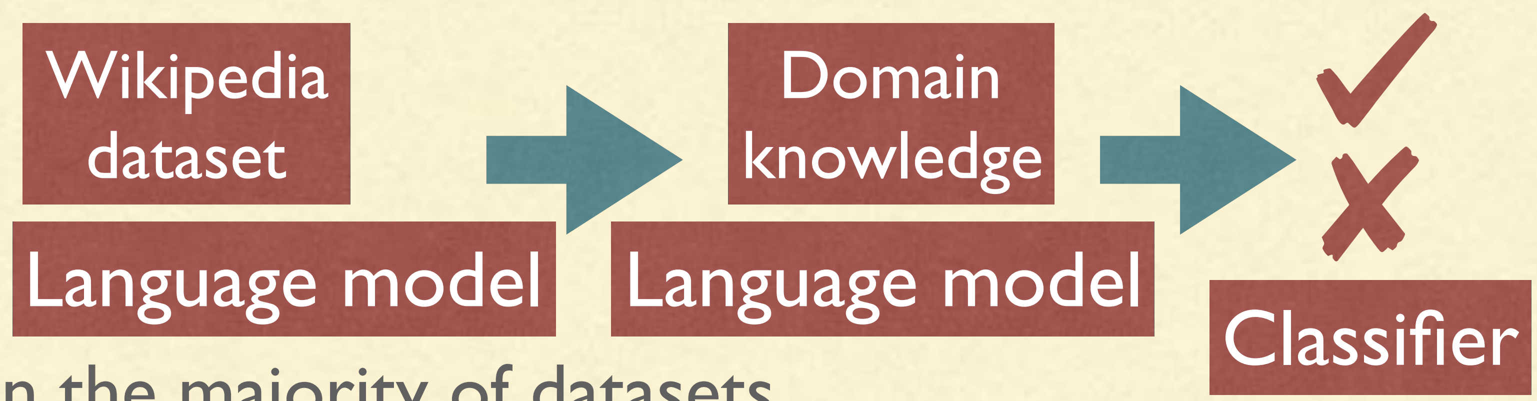


https://arxiv.org/abs/1801.06146

New NLP Approach: ULMFiT

A revolutionary approach for NLP was developed in 2018, called **Universal Language Model Fine-tuning for Text Classification (ULMFiT)**

Using ***neural networks*** and ***transfer learning*** to train a language model, then use it as a classifier. A "***language model***" is any model that learns to predict the next word of a sentence



- can apply it to any task in NLP
- reduced the error by 18-24% on the majority of datasets
- using only 100 labeled examples on 100x more data
- pre-trained models and code are open source

https://arxiv.org/abs/1801.06146

New NLP Approach: ULMFiT

https://arxiv.org/abs/1801.06146

New NLP Approach: ULMFiT

https://arxiv.org/abs/1801.06146

Basic steps:

1. Create (or download a pre-trained) language model trained on a large corpus such as Wikipedia
2. Fine-tune language model using target corpus (such as IMDb movie reviews or hotel reviews)
3. Extract the encoder from this fine tuned language model, and pair it with a classifier. Then fine-tune this model for the final classification task (in this case, sentiment analysis).