Text mining - Sentiment analysis & Knowledge discovery

CSI 5387 Data Mining & Concept Learning

Authors: Xiaoke Liu, Yan Zhang, and Diana Lucaci
October 14, 2020

Table of Content

Problem Definition
Sentiment Analysis

NLP main tasks:

- **Classification**
- Unsupervised **keyword extraction** (important task for Text Mining, and Information retrieval)

Problem Definition
Objectives

1. **Evaluate** the performance of different Machine Learning algorithms on the classification task
2. **Compare** keyword extraction approaches in the context of sentiment classification

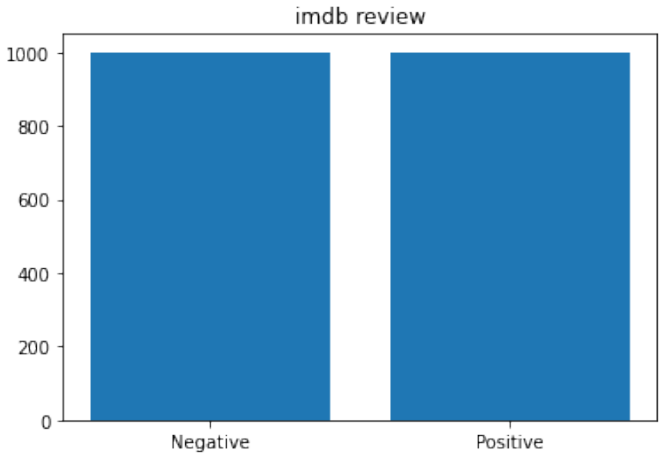
Data Set
Sentiment Labelled Sentences

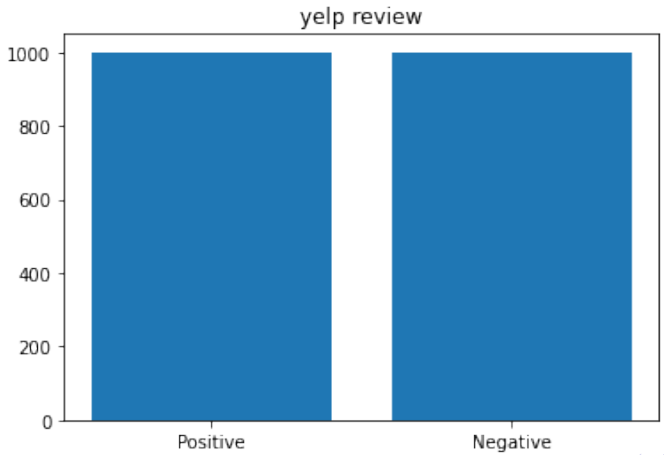
Sentiment Labelled Sentences Data Set From paper 'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015'. The data set contains attributes are text sentences, extracted from reviews of products, movies, and restaurants at different websites. Each records are labelled with positive or negative sentiment with following format:

sentence	score
...	1 (positive)
...	0 (negative)

Table: Format

Data Set
Data Visualization





Data Set
Vectorization

TfidfVectorizer

Convert a collection of raw documents to a matrix of TF-IDF features.

TF: Term Frequency

$$TF(t) = \frac{Number\ of\ items\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

IDF: Inverse Document Frequency

$$IDF(t) = \log_e \frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}$$

Model Construction
Model Selection

To be able to predict new entries, we utilized several methods to build models and fit the given data set. In model construction step, we tried the following options

1. Naïve Bayes
2. SVM
3. MLP

1. **Naive Bayes**

$$P(Class|Sentence) = \frac{P(Sentence|Class)P(Class)}{P(Sentence)}$$

2. **SVM**
   searches for the linear optimal separating hyperplance that best separate different classes.
3. **MLP**
   classify classes based on computational network simulate perceptron

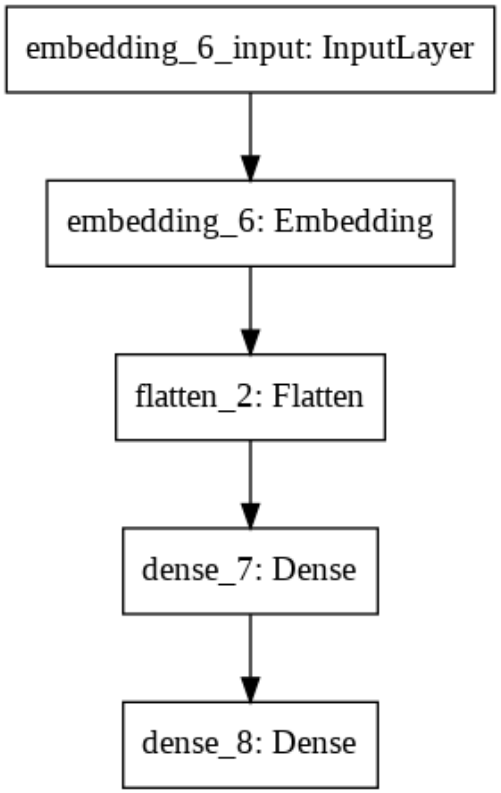


Figure: MLP structure

This simple NN has the structure

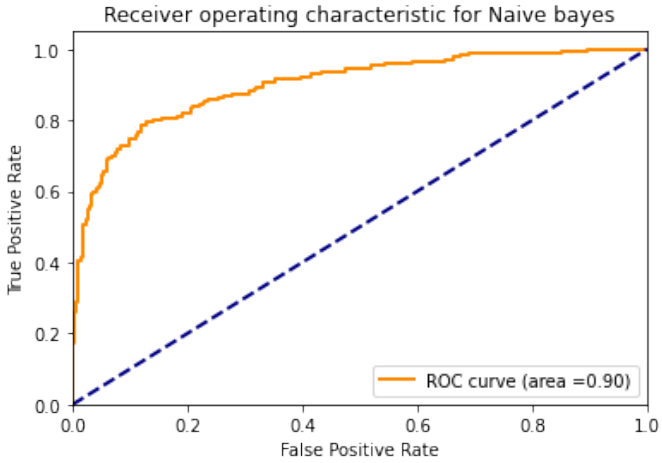
Layer (type)	Output Shape
Embedding	(None, 74, 32)
Flatten	(None, 2368)
Dense	(None, 250)
Dense	(None, 1)

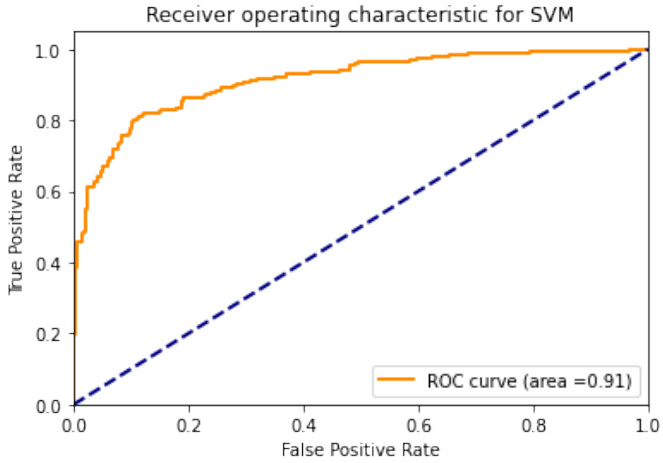
Evaluation - Classification
Result

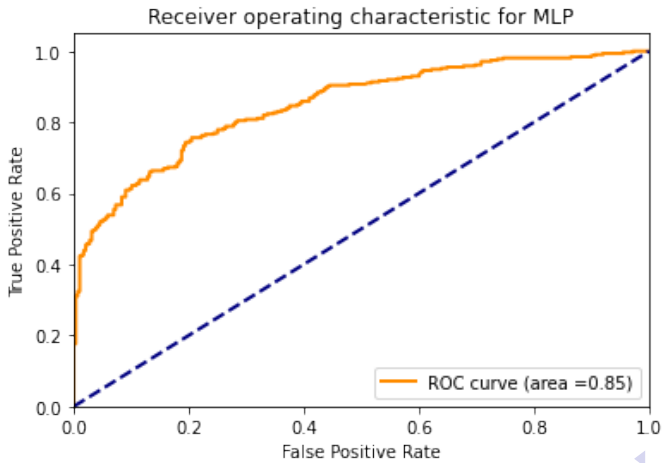
	Accuracy	Sentivity	Specificity	Percision	AUC
Naïve Bayes	82.17%	84%	81%	81%	89.89%
SVM	83.67%	85%	83%	83%	91.19%
MLP	77.33%	76%	79%	80%	85%

Table: Evaluation

Evaluation - Classification
ROC curve







Evaluation - Keyword and Keyphrase extraction
Task and evaluation metrics

Classification

- Accuracy
- Precision
- Recall
- F1
- AUC

Keyword-extraction

- unsupervised task, analyzing the generated phrases/words through:
  - Quantitative analysis
  - Qualitative analysis

Evaluation - Keyword and Keyphrase extraction
Quantitative analysis

- frequency in each corpora
- polarity
- the impact on the classification task

Evaluation - Keyword and Keyphrase extraction
Qualitative analysis

- statistical metrics
- word cloud
- visualization techniques
- application-grounded evaluation

Evaluation - Keyword and Keyphrase extraction Methods

- Statistical methods
  - TF-IDF
  - RAKE
  - YAKE
- Graph-based methods
  - TextRank

Evaluation - Keyword and Keyphrase extraction
Quantitative - classification task

Model	Dictionary	Accuracy	Precision	Recall	F1
Vanilla - LSTM		82.68	0.8449	0.7938	0.8135
LSTM+MLP_gen	TF-IDF	79.44	0.8619	0.6888	0.7612
	RAKE_corpus	82.51	0.8253	0.8158	0.8169
	RAKE_instance	82.68	0.8276	**0.8168**	**0.8186**
	YAKE	**83.00**	0.8501	0.7919	0.8153
	TextRank	82.84	**0.8629**	0.7788	0.8139

Table: Keyword extraction evaluation

Evaluation - Keyword and Keyphrase extraction
Qualitative - metrics

Dictionary type	Avg words / keyphrase (dataset)	Avg words / keyphrase (pos)	Avg words / keyphrase (neg)	Overlap count
TF-IDF	1	1	1	21
RAKE (corpus)	3.61	3.74	3.48	0
RAKE (in-stance)	4.14	4.3	3.98	0
YAKE	1.85	1.89	1.82	23
TextRank	1.13	1.15	1.11	5

Table: Keyphrase analysis for dictionaries of 300 entries

Evaluation - Keyword and Keyphrase extraction
Qualitative

YAKE - positive

YAKE - negative

Contributions
Summary

- Comparative analysis of ML algorithms on the classification task
- Literature review on keyword extraction algorithms
- Evaluation methods for the keyword extraction tasks
- Preliminary results and analysis

Future Work

Possible avenues of extending the projects include

- Apply the methodology on different domains, text data sets, and tasks
- Extending the battery of experiments

Questions and Comments

Thank you.