Yelpilyzer - Sentiment Analysis Based On Yelp Reviews

Power of Natural Language Process

Author: Yisheng Cai, Advisor: Ralph Morelli

Trinity College
HARTFORD CONNECTICUT

Introduction

Natural Language Processing (**NLP**) is one of the newest fields of computer science and it provides quantitative insights into big data.

In this project, the program utilizes **NLP** to train itself over user sentiments behind Yelp reviews on restaurants. A trained program is able to understand reviews and predict a numerical ratings for future reviews.

Approach

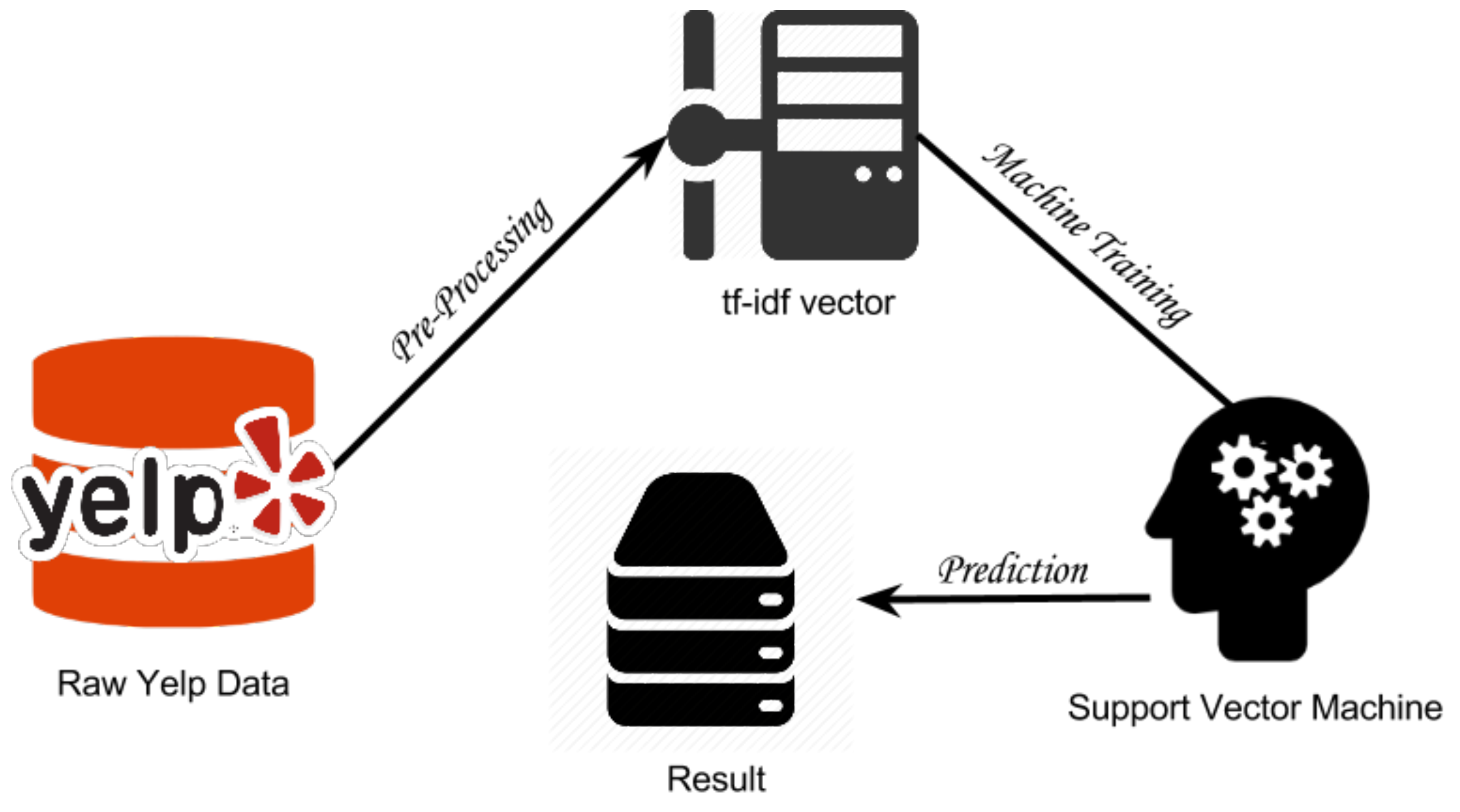

Figure-1 Procedure of Yelpilyzer

The project takes a collection of Yelp reviews on businesses, and performs following tasks:

- Pre-Processing by **Vectorization** and **Tokenization**
- Machine Training using **Support Vector Machine**
- Prediction: Predict sentiment of new input reviews

Vectorization

The key of vectorization is to help the machine understand the meaning of words, quantitatively, for which, a **tf-idf** conversion is introduced, as shown in Figure-2.

Term Frequency-inverse Document Frequency (tf-idf), is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The corpus, in this case, is the set of all Yelp reviews.

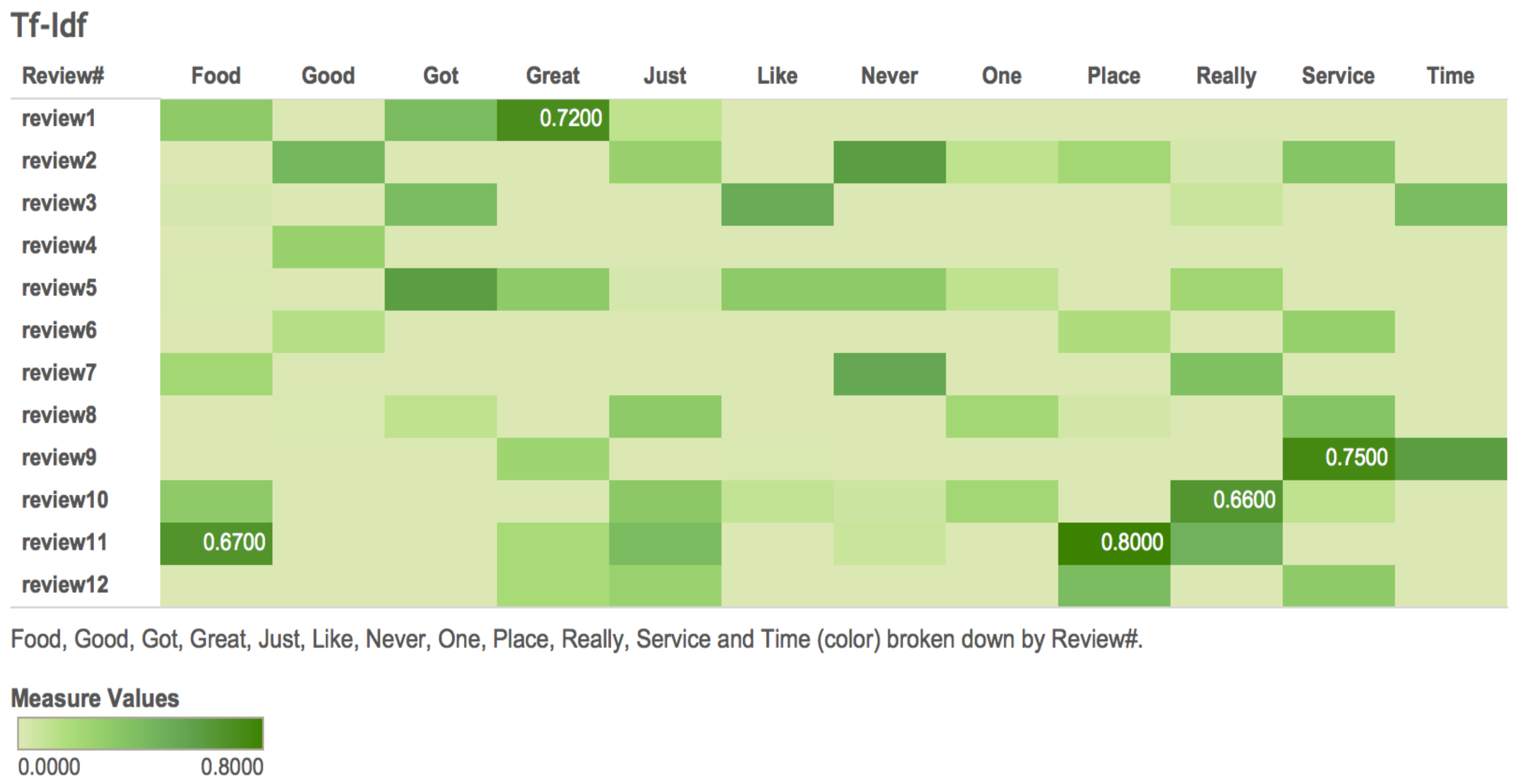

Figure2 Sample **tfidf** based on a 12-word matrix

Tokenization

The program tokenizes reviews and removes meaningless English words from the input vector to improve accuracy. Some most frequent words are displayed below in Figure-3:

good
great
food
place
service
one
like
restaurant
ordered
pizza
menu
beer
cheese
really
chicken
bar
time
back
salad
better
just
nice
best
little
try
go
friendly
delicious
Chicago
table
going
made
know
definitely
experience
sauce
want
night
even
dishes
staff
wait
get
half
well
also
always
Rachel
much
meal
never
drinks
dining
small
eat
sushi
sure
gluten
order
wine
asked
dinner
anything
breakfast
bit
probably
took
make
fresh
sandwich
something
lunch
went
free
come
people
area
came
pretty
two
dessert
give
perfect
tasty
steak
staff
server
recommend
cheap
Greek
bagel
pasta
way

Figure-2 Word-Cloud generated using most frequently appeared words in Yelp reviews

Support Vector Machine

The machine learning model of this project is a Support Vector Classifier/Machine.The algorithm partitions data into five classes, 1-star (most negative) to 5-stars (most positive) respectively. The advantages of using SVC are:

- ❑ Effective on high dimensional spaces and when number of dimensions is greater than number of samples
- ❑ Versatile because different Kernel functions can be specified for the decision function

Multiple kernels for this partitioning algorithm are shown in **Figure-4**. After a few iterations, linear kernel out-performs other kernels on accuracy.

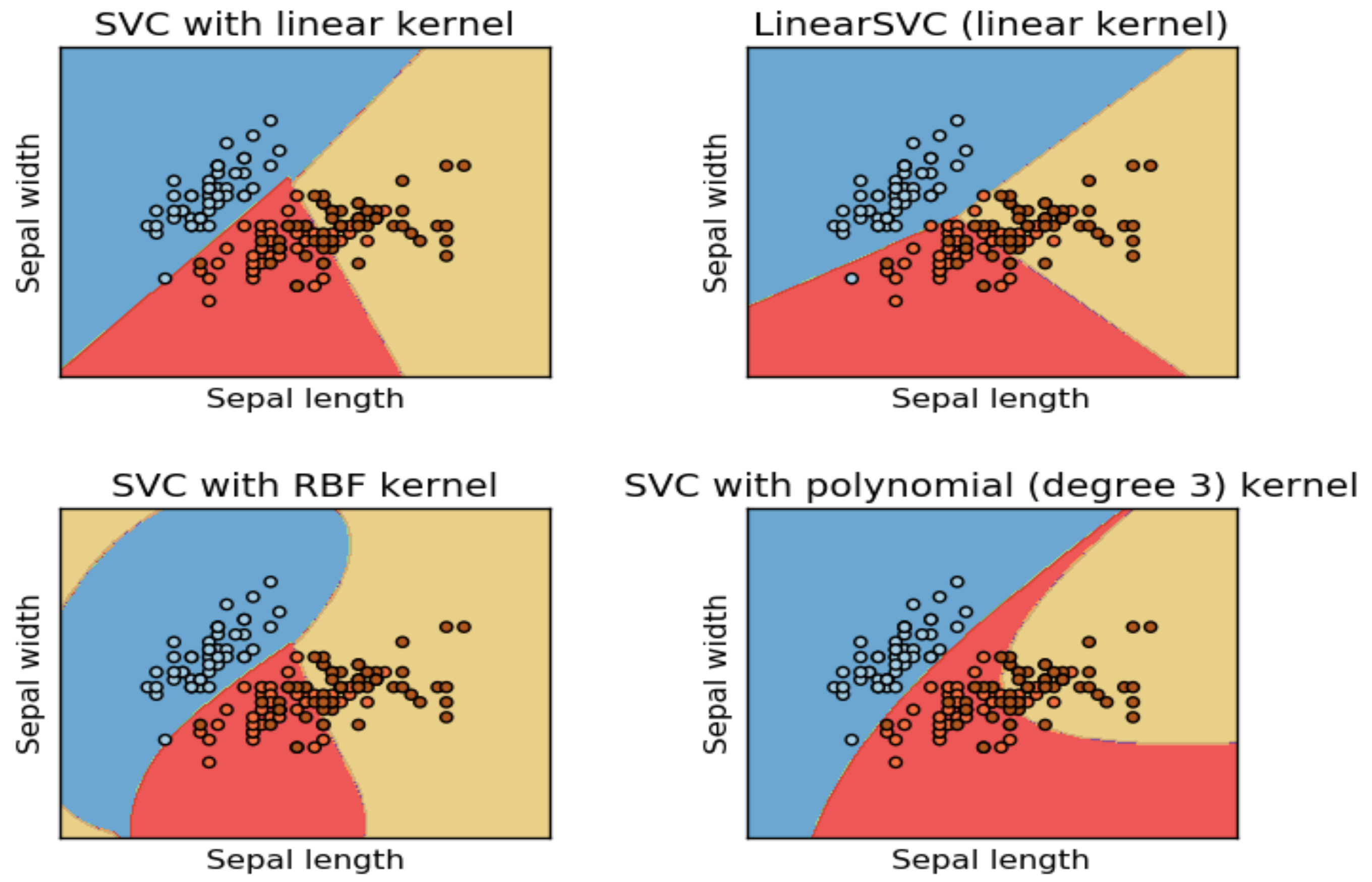

Figure-4 Various kernels available for SVC

Results

The result table below shows promising accuracy on training and predicting sentiment over 1,000,000 reviews.

```
========================    Sample Tf-Idf Vector    ===========================
 [ "ask"       "authentic" "burger"    ...,  "yummmm"    "zero"      "zucchini"]
[[ 0.          0.          0.          ...,  0.03841453  0.          0.        ]
 [ 0.          0.          0.          ...,  0.          0.          0.        ]
 [ 0.06113526  0.          0.          ...,  0.          0.          0.        ]
 ...,
 [ 0.          0.          0.26048377 ...,  0.          0.          0.        ]
 [ 0.00319634  0.          0.          ...,  0.44341896  0.          0.        ]
 [ 0.          0.          0.          ...,  0.          0.          0.        ]]

1000000 Rows
18623   Columns
========================     Prediction Sample    ============================
prediction: [4 3 4 5 3 5 1 4 4 5 2 2 2 2 4 3 4 4 5 1 5 1 4 5 1 2 3 1 2 5 4 4 1]
target:     [3 4 4 5 3 4 3 4 3 2 3 2 4 2 5 5 2 4 3 1 4 1 2 5 1 2 4 1 3 5 3 4 2]
R-square: 0.837
train accuracy: 0.984
===============================================================================
```