CS584 – MACHINE LEARNING

FALL 2016

Sentiment Analysis

Group Members:

Anup Deulgaonkar

Subhadeep Roy

Table of Contents

Task

Dataset

- Data source
- Target variable
- Features
- Data size

Preprocessing

Visualization

- Target
- Features

Evaluation

- Performance Measure
- Classifiers
- Evaluation Strategy
- Performance Results
- Top Features
- Discussion

Interesting/Unexpected Results

Contributions of Each Group Member

Conclusion

References

Sentiment Analysis for Movie Reviews

Group Members: Anup Deulgaonkar, Subhadeep Roy & Yogesh Palav

Task

Classification of Movie Reviews as Positive or Negative by performing Sentiment Analysis. Sentiment Analysis on social media can help understand generic opinion and identify KPI's and improve them.

Dataset

The corpus consists of set of files with reviews from users. There are 25000 reviews which are equally divided and labeled into positive and negative reviews.

Data source

We got the data from IMDB. We did not label any of the reviews manually.

Target variable

Categorize the given review as Positive/Negative (Target Variable).

Features

Input Feature is a Bag-of-Words Representation which has the vocabulary of size 47460 represented in a sparse matrix.

Data size

25000 Instances of Positive and Negative with each of them containing 12500 instances.

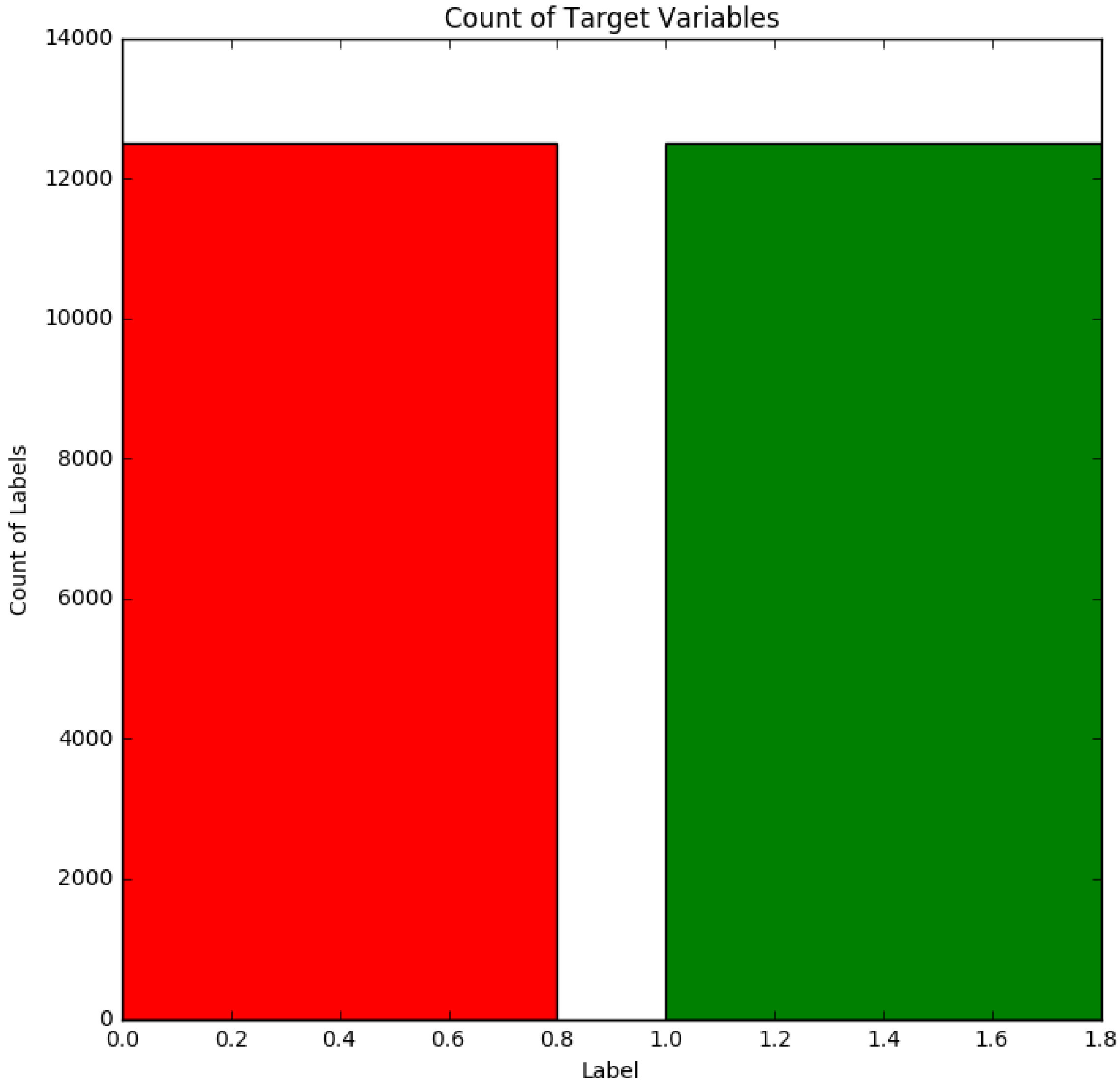
Preprocessing

We have a pipeline kind of a mechanism in which sentences are first Stemmed, Lemmatized and convert to Bag-of-Words using TFIDF Vectorizer and are normalized by lengths.

Visualization

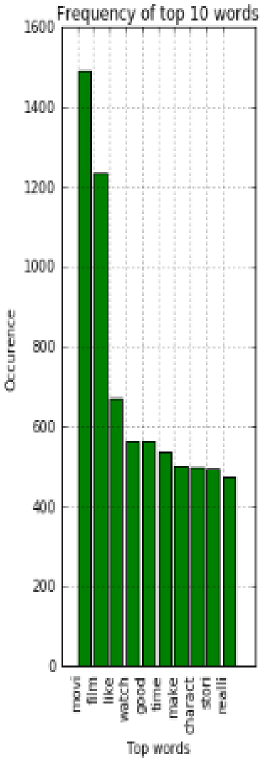
Target

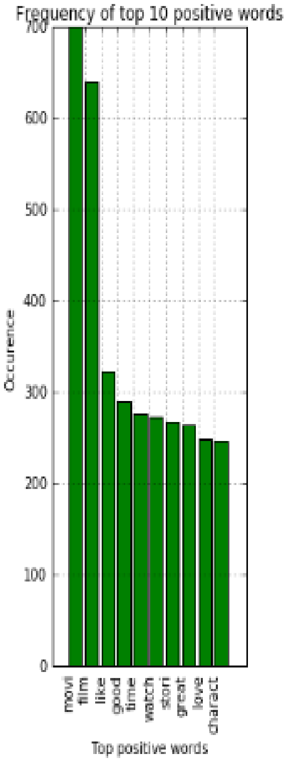
12500 Positive and 12500 Negative Instances

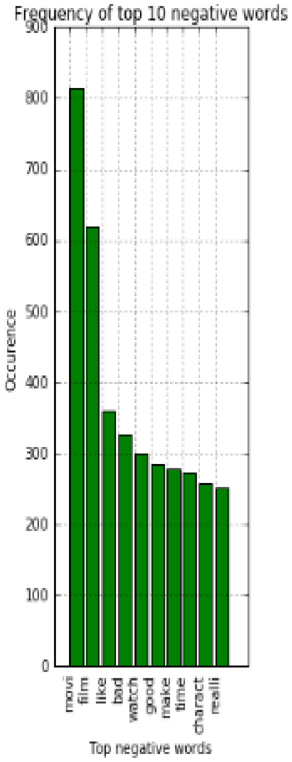


Features

Top Frequency (Hig TFIDF weights) of words in each category.







Evaluation

Performance Measure

Accuracy cannot be used as a performance criterion because classifying everything as positive or negative gives 50% accuracy and the model is useless. Hence, we used precision as the performance criterion.

Classifiers

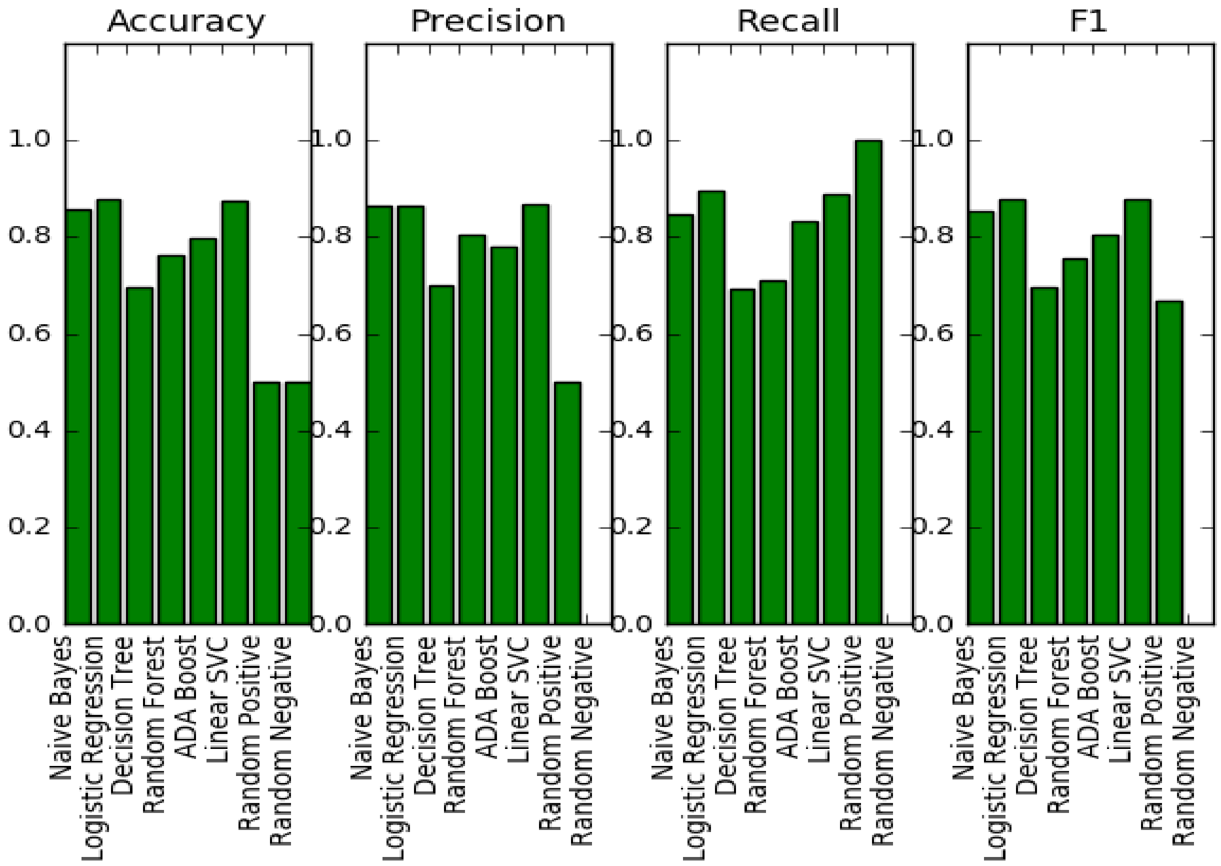
Model	Parameters	Performance (Precision)
Multinomial NB (Baseline)	alpha=1.0, fit_prior=True	0.86
Logistic Regression	penalty=l2,C=1.0,fit_intercept=True	0.89
Decision Tree	random_state=0,criterion=”gini”	0.71
Random Forest	criterion=’entropy’	0.77
Liner SVC	penalty=’l2’,C=1.0,fit_intercept=True	0.88
ADA Boost	random_state=none	0.81

Best model and parameter setting was ->Logistic Regression with a precision of 0.89

Evaluation Strategy

we performed a train-test split of 0.67 and 10 fold Cross-Validation.

Performance Results(Baseline: Naive Bayes)



Top Features

According to the performance criterion (precision) Logistic Regression outperformed other models.

The top features of LR are:

1. Positive Class:
    a. great, excel,perfect,love,best, enjoy,favorit,beauti,fun,amaz
2. Negative Class
    a. worst,bad,wast,poor,bore,noth,disappoint,wor,horribl,suppo,

Discussion

W used Bag-of-Words representation which gave us okay results but when performed some analysis on the wrong instance classification the main problem was the context of the word. For Example: some noun/ noun forms (like actress kareena kapoor) were having high weights for negative which on given a simple sentence with that keyword would categorize the sentence as negative. A much better representation would be to be perform POS Tagging on the data and normalize the weights of the POS which actually are used to convey the sentiment of the sentence then the ones which do not.

Interesting/Unexpected Results

1. The model uses Bag-of-Words so in the previous slides some of the noun forms have higher negative weights.
2. For Example: A simple review 'movie was okay. The woman in red acted nice' was categorized as negative.
3. The problem is with context in which the word is used and one of the solution is to either use N-Grams or N-order Markov Model.

Contributions of Each Group Member

Name	Task
Anup Deulgaonkar	W used Bag-of-Words representation which gave us okay results but when performed some analysis on the wrong instance classification the main problem was the context of the word. For Example: some noun/ noun forms (like actress kareena kapoor) were having high weights for negative which on given a simple sentence with that keyword would categorize the sentence as negative.
Subhadeep Roy	W used Bag-of-Words representation which gave us okay results but when performed some analysis on the wrong instance classification the main problem was the context of the word. For Example: some noun/ noun forms (like actress kareena kapoor) were having high weights for negative which on given a simple sentence with that keyword would categorize the sentence as negative.
Yogesh Palav	W used Bag-of-Words representation which gave us okay results but when performed some analysis on the wrong instance classification the main problem was the context of the word. For Example: some noun/ noun forms (like actress kareena kapoor) were having high weights for negative which on given a simple sentence with that keyword would categorize the sentence as negative.

Conclusion

References

http://radimrehurek.com/data_science_python/

https://www.cs.tut.fi/sgn/arg/klap/introduction-**semantics**.pdf

https://www.nltk.org (Chapters on stemming and tokenizing)