Sentiment Analysis of IMDB Movie Reviews

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Team Members -

SPRING 2020
CS-580L-01

- **Chaitanya Kulkarni**
  B-Num: B00814455
  SUNY Binghamton
  ckulkar2@binghamton.edu

- **Abhimanyu Singh**
  B-Num: B00813542
  SUNY Binghamton
  asing134@binghamton.edu

- **Necati Anil Ayan**
  B-Num: B00777933
  SUNY Binghamton
  nayan1@binghamton.edu

1. Motivation

- Most convenient source of entertainment
- Often confused about whether to watch that particular movie or not
- We check websites for ratings and reviews
- Most of them show rating based on the stars given
- But no method to know about the success of the movies based on the reviews and comments

2. Introduction

- To determine the success or failure based on the reviews, we analyze the text
- Interpret and classify the emotions within textual data
- Focuses on polarity - Positive reviews or Negative reviews
- We used 3 classifiers -
  - Logistic Regression
  - Naive Bayes
  - Support Vector Machine (SVM)

3. Dataset Info

- **The dataset can be found at -**
  Large Movie Review Dataset v1.0

- **Dataset provided by Stanford**
- **50,000 movie reviews from IMDB website**
- **25K for training data & 25K for testing data**

4. Data Preprocessing

- Removing HTML tags
- Remove special characters and stopwords
- Lemmatization
- Tokenization

5. Feature Extraction

- **Is used to convert feature (words in this case) into some form.**

Bag of Words Approach

- Term Frequency
- CountVectorizer with scikit-learn

6. Classification Models

- Classifiers -
  - Logistic Regression
  - Naive Bayes
  - Support Vector Machine (SVM)

A. Logistic Regression

- Discriminative Model
- Logistic Function (Mostly Sigmoid)
- Pretty good for binary output model (positive or negative)
- Not good for non-linear solutions

B. Naive Bayes

- Generative Model
- Assumes all features are conditionally independent
- Based on Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- May be absurd for real-life cases since predictors are usually dependent.

C. Support Vector Machine (SVM)

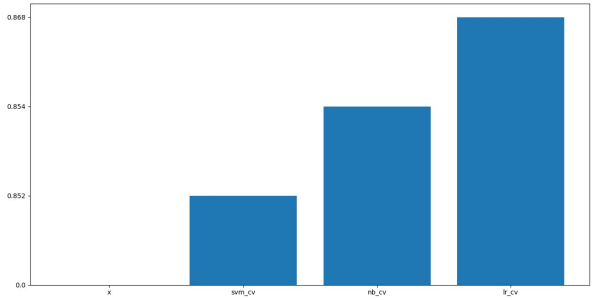
- Objective is to define hyperplane in N-dimensional space to classify data points.
- Goal is the maximizing margin (between two data points)
- Also good for non-linear solutions (Kernel trick)
- Very suitable for text classification

7. Result Evaluation

- **ACCURACY**
  - A result evaluation measure
  - It is the ratio of correctly predicted observations to the total number of observations
  - Higher the accuracy better the model

7. Result Evaluation(cont.)

- **LOGISTIC REGRESSION**
  - Accuracy: 86.89%
- **SUPPORT VECTOR MACHINE**
  - Accuracy: 85.29%
- **NAIVE BAYES**
  - Accuracy: 85.48%



8. Conclusion

- Highest accuracy obtained on Logistic Regression model.
- LR performs better than Naive Bayes
- LR outperforms the SVM, which is known to be the best choice for textual data.

9. Our Learning

- Before handling the data we need to clean the data. This is called data preprocessing. Necessary to clear missing, noisy and inconsistent data which might affect accuracy.
- Semintent Analysis is extraction of the emotions from within the textual data. It is a technique based upon Natural Language Processing.
- Logistic regression and support vector machines are closely linked. Both can be viewed as taking a probabilistic model and minimizing some cost associated with misclassification based on the likelihood ratio.
- Naive Bayes classifier is the generative model. Naive Bayes also assumes that the features are conditionally independent.
- Both Naive Bayes and Logistic regression are linear classifiers, Logistic Regression makes a prediction for the probability using a direct functional form whereas Naive Bayes figures out how the data was generated given the results.
- Thus, when the training size reaches infinity then discriminative model logistic regression performs better than the generative model Naive Bayes.
- Hence in our project we get more accuracy for Logistic Regression than the other two classifiers.

- Chaitanya Kulkarni

9. Our Learning

- Understood that main difficulty is to make dataset ready for training (Data Preprocessing).
- Understood the differences between generative and discriminative models.
- When dataset size is very large like infinite discriminative models are preferable. However, Generative models can reach its asymptotic faster since it needs fewer training set to do that.
- As known that SVM is the one the best classifiers. However, it performs poorly in our dataset. Main reason of this may be we have ignored hyperparameter tuning. If we consider hyperparameter tuning, accuracies for SVM and LR may increase but Naive Bayes accuracy may remain stable.

- Necati A Ayan

9. Our Learning

- Got to learn that, constructing and training a model is only a small part of a machine learning project, major task lies in Data Pre-processing
- Learned various ways to clean the textual data and why is it necessary, to make it suitable to feed to our learning model.
- Learned to apply what was taught in class in our project, like which models to choose for a specific task, which in our case was text classification.
- Learned how to improve the accuracies of the models by employing various optimization techniques like hyperparameter tuning
- Learned the actual difference between probabilistic and Binary Classifiers
- Got to know that, though some models are considered better for some tasks, but other models can outperform them in some data-sets, which happened in our case.

- Abhimanyu Singh

10. References

[1]. MaisYasen, Sara Tedmori. “Movies Reviews Sentiment Analysis and Classification”. IEEE Jordon International Joint Conference on Electrical Engineering and Information Technology (JEEIT). 978-1-5386-7942-5.

[2]. Tirath Prasad Sahu, Sanjeev Ahuja. “Sentiment Analysis of movie reviews: A study on feature selection and classification algorithms”. International Conference on Microelectronics, Computing, and Communication (MicroCom).978-1-4673-6621-2.

[3]. Wijayanto**,** Unggul and Sarno, Ritanarto. “An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naïve Bayes”. 476-481.10.1109/ISEMANTIC.2018.8549788.

[4]. Tejaswini M. Untawale, G. Choudhari. “Implementation of Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches”. 978-1-5386-7808-4.

Thank You