Miniproject 2: Sentiment Analysis from IMDB Movie Reviews by Using Different Classification Models

Ali, Mumu Aktar

May 26, 2019

1 Abstract

In this project, we focused on developing different machine learning models to predict the IMDB movie reviews as positive or negative sentiments and finding the best performed model with the most significant feature set based on their classification accuracy. Significant feature extraction i.e., TF-IDF, binary occurrences and N-grams are applied on the text data after cleaning up the data by removing punctuation marks, stop words, transforming links as well as url to words, tokenizing, lemmatizing and taking the lowercase of the text. Bernoulli Naive Bayesian with Laplacian smoothing is performed taking the binary word occurrences as features which gave 89% of accuracy. Further, four different classification models: logistic regression, random forest, support vector machine and deep neural network are developed from Scikit learn considering continuous features to analyze the positive and negative sentiments of movie reviews. Best feature set is obtained by varying the maximum as well as minimum document frequency and N value for N-grams which affect the classification accuracy significantly. A model validation pipeline is developed using 10-fold cross validation for the models used and performance of logistic regression using binary occurrences as well as TF-IDF features are analyzed. Finally, with the best performing feature set we applied the other models on the dataset among which deep neural network gave the best of 91% of accuracy in classifying positive and negative reviews.

2 Introduction

IMDB is a popular website for movie information and reviews where both positive and negative comments come from viewers for the movies. For positive and negative sentiment analysis based on the reviews and comments, Bernouli naive bayesian, logistic regression, random forest, support vector machine and finally deep neural network are applied on the dataset to analyze the performance of these models as well as to find the best performing model. Accuracy and runtime are used as the two performance analysis tools for the models. Since the performance varies based on the features used, significant feature extraction is performed on the training and testing set after cleaning up the messy data.

CountVectorizer is used to create a bag of words (BoW) considering lemmatization to recover the canonical form of words, maximum document frequency, max_df to avoid words occurred very frequently and don't contribute much to the classification, minimum document frequency, min_df to consider a minimum number of documents a word should be present to be considered as a feature and N-grams to take the paired words. These are tuned here to obtain an optimal size of vocabulary as the larger vocabulary set makes more sparse representation and reduced small set ignores most important and frequent words necessary to analyze the reviews. To extract significant text features along with the binary occurrences obtained from BoW, TF-IDF is calculated for obtaining the indicative words as well as their relative frequency in the overall documents. The parameter tuning, cross-validation with grid search and varying performance of TF-IDF and N-grams are performed on logistic regression model implemented using pipeline through Scikit learn. Also other models are applied on the best feature set using the same pipeline and 10-fold cross validation to analyze their performances. Bernoulli Naive Bayesian with Laplacian smoothing is implemented (from scratch) to overcome the effect of zero probability that occurs when an unknown word comes in testing which is not from the given vocabulary. In that case, rather than likelihood of 1 it gives 1/2 which helps to remove the undefined problem with log being used. Finally, best accuracy of ... is obtained by ... using the best feature set to classify the positive and negative movie reviews.need to write the best model with accuracy........

3 Related Work

Different literatures have considered NLTK and Scikit Learn to extract features for sentiment analysis in movie reviews. To understand the basic of vectorization,TF-IDF as well as vocabulary creation using Scikit learn, we used [1] as reference where SVM and multinomial naive bayesian are applied to analyze the sentiment alalysis after feature extracion. Similar work has been done using multinomial naive bayesian in [2] where Bag of Words (BoW) has been used to extract features from the data collecetd from a movie site and to fit the linear classifier, logistic regression is applied to compute r and b coefficients. A number of different heterogenous features: SentiWordNet Score, SentiWordNet Score for adjective only,term presence of SentiWordNet Score, top 10 TF-IDF SentiWordNet, exaggerated word shortening, exaggerated word shortening, intensifier handling and emoticons only, etc. have been extracted for classifying sentiment class level as positive or negative using linear SVM and Naive Bayesian with higher accuracy by Naive Bayesian [3]. In the recent era, deep learning is also showing significant performance in sentiment analysis task which is chosen also in our work.

4 Dataset

Dataset is collected from kaggle for IMDB movie reviews and preprocessing is performed on the text data by taking lowercase, removing the stop words, cleaning up the punctuation marks, transforming the https or url link as a word 'url' and @ as 'at_user' as those do not affect the classification task significantly.

Cleaning up as well as tokenizing the texts are performed in the same manner to make numeric representation of them to fit as inputs into all machine learning models. Before performing feature extraction, train-test splitting is performed using Scikit learn with test size=0.10 to consider more training data for 1-fold cross validation and further test size=.30 for the 10-fold cross validation and random state=42. Finally continuous data are normalized to rescale the varying attributes for both training and testing to a unit form before fitting them into the classifier models.

5 proposed methodology

To predict the positive or negative review of movies, among all the developed models, logistic regression is chosen to tune the parameters for extracting significant features with the best combination of values of max_df,min_df and N-grams. Logistic regression is a probabilistic approach which maximizes the likelihood of features by directly modeling the log-odds on linear function to classify binary targets. Experiment with logistic regression is performed using Scikit Learn [l] with L2 regularization as a penalty to the cost function. Two separate feature extractions: binary occurrences obtained from the BoW and TF-IDF weighting applied on the text features are performed on the train and test data separately and their performance is compared for the best parameter setting. 10-fold cross-validation is performed with grid search using different values of C: [0.01, 0.1, 1 ,2, 3, 4, 6, 8] where best value of C varied for different parameter setting and max_df=0.95,min_df=5, and N-gram=2 gave 91.12% of accuracy for test_size=10% and 90.28% for test_size=30%. To select the best parameter settings, a number of different 1-fold cross-validation with different combination of max_df,min_df and N-grams are analyzed at first and their performance is shown in the curves (Fig.1 and Fig.2) below. Due to the time complexity, 1-fold cross validation is performed at first for finding better combination of max_df,min_df and N-gram. Finally, with the best combinations, four more operations are performed with 10-fold cross validation whose results are highlighted in the result section. Finally, binary occurrence is considered with this best setting of vectorizer(max_df=.95,min_df=5 and N-gram=2) and 10-fold cross validation is performed with grid search which gave 89.68% of accuracy with best C=0.1 and runtime 2403.86 s. After finding the best text feature set, support vector machine, random forest and deep learning are applied on these features and 10-fold cross validation is performed for each model. ...need to write about best performing model...

From the above curve, it can be seen that for 1-fold cross validation, we've got the best accuracy of 0.912 for logistic regression with max_df=.95, min_df=5 and N=2 and quite similar accuracy of .9112 with max_df=.995, min_df=30/35 (same result) and N=3. It is noticed that the ratio of max_df and min_df should be an optimal one. It is not the case that we can select the highest value of them to get the best feature set rather we can see that max_df=.95 which eliminates more words to be counted as features considers min_df=5 than max_df=.995 which needs min_df=30 to work as the best parameters. Although N=3 gave better results for than N=1 but N=2 gave the best result with a good combination of max_df and min_df which means that incresing values of paired up words doesn't affect significantly to the best feature extraction rather

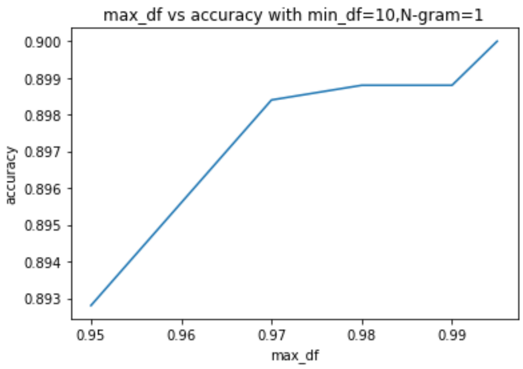

Figure 1: max_df vs accuracy with Ngram=1 and min_df=10

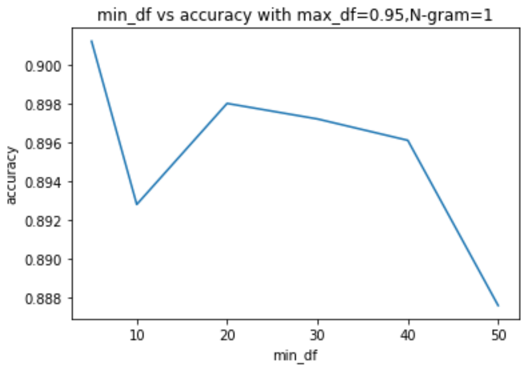

Figure 2: min_df vs accuracy with Ngram=1 and max_df=95

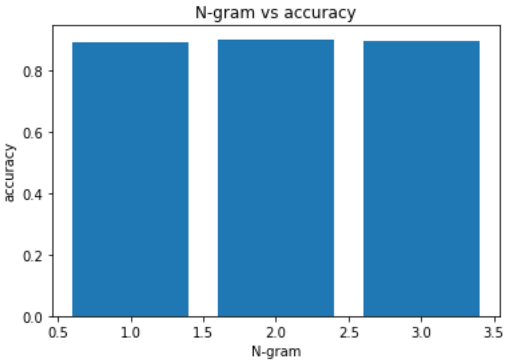

Figure 3: N-gram vs accuracy for different values of max_df and min_df

we found N=3 for max_df=.95, min_df=5 considers a lot of features which we couldn't utilize because of memory error.

6 Results

Table 1. shows the performance of logistic regression for four different combination of max_df, min_df and N-gram which are the best results obtained from the previous 1-fold cross validation done with 10% of test data. The runtime is based on the memory occupied, so it can vary although it took more time for N-gram=2.

max_df	min_df	N-gram	Accuracy	Runtime	Parameter,C
.995	10	1	88.90	242.33 s	4
.95	5	1	89.24	316.844 s	6
.95	5	2	90.28	6318.57 s	8
.995	30	3	89.6	488.229 s	8

References

[1] Abhijeet Kumar, *https://appliedmachinelearning.blog/2017/02/12/sentiment-analysis-using-tf-idf-weighting-pythonscikit-learn/*

[2] Martín Pellarolo, *https://medium.com/@martinpella/naive-bayes-for-sentiment-analysis-49b37db18bf8*

[3] Rachana Bandana, *Sentiment Analysis of Movie Reviews Using Heterogeneous Features* 2018