TED Talks - A Predictive Analysis Using Classification Algorithms

Kumkum Yadav
MSIM, School of Information Sciences
University of Illinois, Urbana Champaign kumkumy2@illinois.edu

Garima Garg
MSIM, School of Information Sciences
University of Illinois, Urbana Champaign
garimag2@illinois.edu

***Abstract*:** TED is the acronym for Technology Entertainment and Design.The talks in the platform are a great source of knowledge and ideas which are available online for free. TED talks are held on a plethora of topics such as Technology, Entertainment, Design, Cultural, Academic Researches etc which are presented by distinguished speakers. The purpose of this study is to predict the number of views of the talk and analysing the overall reaction of the talks based on the user comments.The dataset is taken from www.kaggle.com and it include details of 2550 TED Talks from 2006 to 2017. There are 3 main goals for this paper. Firstly, to predict the number of views for TED talks based on different attributed given in the dataset. Secondly, to predict the sentiments of the ratings for TED talks based on the description of the comments by the users. Thirdly, to get some visualizations done on the dataset to get more understanding about the topic and the results which we get. In addition to this, the paper also discusses the observations and the limitations on the project.

***Keywords*: classification, data mining, prediction algorithms, visualization, Naive Bayes, KNN, SVM, Linear Regression, confusion matrix, precision, recall.**

I.INTRODUCTION

The first thing which comes to our mind after listening to the word 'TED Talks' is "Amazing Ideas". TED is an organization which is devoted in spreading ideas to change attitudes, lives and ultimately the world. TED with slogan "Ideas worth spreading", started in 1984 in Silicon Valley. Earlier, the idea was to include talks only on Technology, Entertainment and Design (hence the name) but later on with growing popularity and demand the topics were broadened to include cultural, societal and other academic topics. The main essence of TED talks is to democratize knowledge. TED Talks are full source of knowledge and the valuable insights from the talks are available free online. Now there are more than 200 talks per year. The talks are varied like short movies presented by eclectic speakers who have a deep professional knowledge and experience about the topic they choose to talk about. The talks are focused, inspiring and popular which goes beyond classroom and office. There is a great dispersion in the number of views of the talks which made us eager to work on the dataset related to TED Talks.

II. TOOLS & SOFTWARE USED

The following are the tools and softwares used for the project:

- OpenRefine - We used the tool to clean the data.
- Excel - We used Excel for the data management and for streamlining of the data
- Python, Weka - We used the Python platform for applying the classification algorithms and some of the visualizations.
- Tableau - We used Tableau to understand the data set, and split some of the columns to get more visualizations.

III. DATASET

We took the dataset from www.kaggle.com, which has the details of around 2550 TED talks from year 2006 till 2017. The dataset also contains information about all audio-video recordings of the talks. In the dataset, there are columns which describes the characteristics of the talks such as the title, description, transcript, speakers, duration etc. and there are other columns which tells about the columns which has impact on the views of the talks like comments, languages, and views. The main data set consists of following important columns:

- Name – The name of the TED Talk
- Title – The title of the talk
- Description – Description for publicity of the talks
- Main_speaker – The name of the speaker of the talk
- Speaker _occupation – The occupation of the main speaker
- Num_speaker – The number of speakers of the talk
- Duration – The duration of each talk in seconds
- Event – The TEDX event when the talk took place
- Film_Date – The date when the talks took place
- Published_Date - The date when the talk was published
- Comments – The number of comments for the talks
- Tags – The description of themes associated with the talks
- Languages – The number of languages in which the talk is available
- Ratings - The combination of various ratings given to the talks by various users

· Related_talks – A list of recommended talks which can be watched next
· URL – The URL of the talks
· Views – The number of views for each talk

From the above dataset we designed two predictive models keeping “Views” as the class attribute in our first model and “Ratings” as the class attribute in the second model.

IV. DATA CLEANING AND PRE-PROCESSING

The dataset from Kaggle was already cleaned except for few special characters in the ‘Transcript’ and the ‘Description’ column. The special characters in the columns like (; ‘ “ | * ã á) were removed using the tool , OpenRefine.

As a preprocessing step,we needed to correct the dates in the data set as they were not in correct date format. The dates were displayed in some numeric format which was hard to understand. We changed the date to the correct format using python code.

In this study, we first predicted the number of views but there was huge dispersion in the number of views and the data was highly skewed. So, on the basis of median, we divided data in 2 parts. In the first part the views were above median and in the second part the views were below median. We removed all the outliers from the data and ran test on the new data to get better results.

We also divided the data in binary form. 1 (high views) for numbers which are more than median and -1 (low views) for numbers which are less than median. By doing this, we could predict which talks got high views and which got low views. For the second prediction where we analysed the sentiments of the ratings ,we considered the ‘Ratings’ column which had ratings with many words or phrases in it. We used tableau to split the words in different columns. After that we applied filters and calculated the frequency of words. Then we manually categorized those words in positive, negative and neutral ratings. Words like *confusing, obnoxious, unconvincing, long winded* were taken as negative words and words like *funny, beautiful, persuasive, ingenious* were considered positive words and *ok* as the neutral word.

V. VISUALIZATIONS

We used Tableau to come up with some interesting visualizations.

Number of talks published per year

Fig. 1 shows that the number of talks published per year gradually increased from year 2006 to year 2012 and after 2012 there is a sudden and slight drop in the number which was consistent till 2017. After 2012, we notice that the average number of talks per year is around 200.

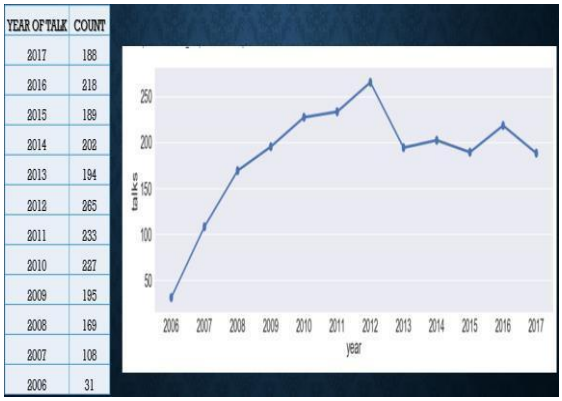


Fig 1. Number of views on talks per year

Number of views on talks per year

The number of views on the talks per year (published date) is shown in the below graph fig 2. The graph shows that the number of views increased gradually from year 2006 till 2016. But we can see that after 2016, but it decreased remarkably from 2016 to 2017. This drastic change in the numbers could be because the talks published in 2017 are very new and did not get enough click throughs. Also the talks might not get that popular so soon.

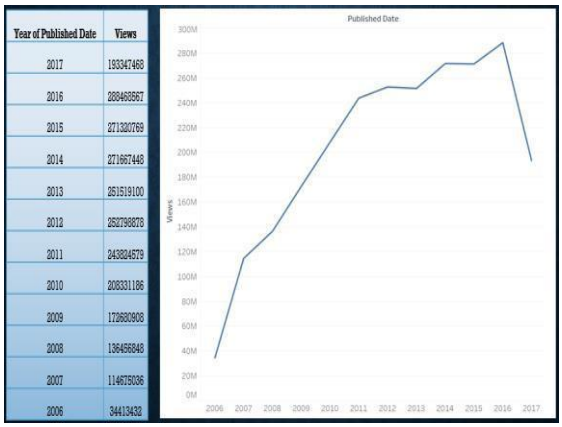


Fig 2. Number of views on talks per year

Top 10 Names of Speakers Based on Number of Views

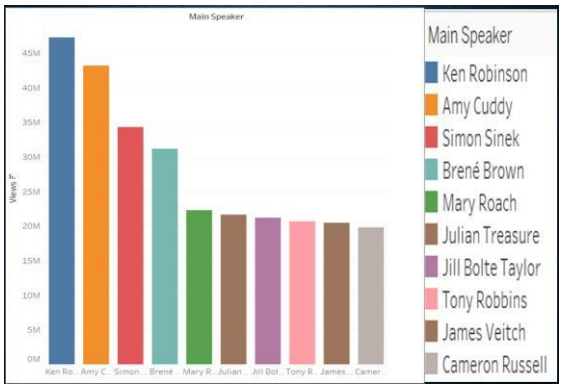


Fig 3. Top 10 names of speakers from TED Talks

The above graph fig. 3 shows the top names of speakers from the TED Talks who got maximum number of views. These are the most popular speakers. The speakers are have a great knowledge in their respective fields. This is also one of the factors which predicts the number of views or the popularity of the talks. When we looked for the occupation of the top speakers, we got to know that most of them are known authors.

Occupation of speakers with maximum views

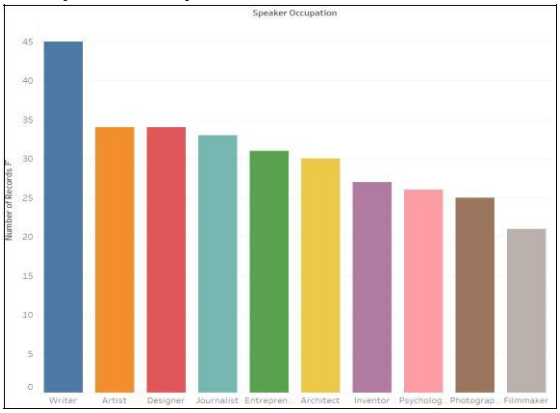


Fig 4. Occupation of top speakers

The above graph fig. 4 shows the occupation of the top speakers with the maximum number of views. One interesting thing noticed here is that among the top speakers, 'Architect' are also present who gave the talks and it was popular. Another noticeable thing is that the author or the writers got the maximum number of views.

VI. CLASSIFICATION ALGORITHMS

The following classification algorithms were applied to the data set :

1. Logistic Regression
2. Support Vector Machine
3. K Nearest Neighbour
4. Naive Bayes

Prediction 1: Predicting the number of views based on number of speakers, languages,duration and number of comments.

Class Attribute: View Type

Predictors: Languages, Comments, Duration, Rated As, Number of Speakers

Prediction 2: Predicting the overall sentiments of the ratings based on the description of the comments by the users.

Class Attribute: Rated As

Predictors: View Type, Languages, Comments, Duration, Number of Speakers

For the first prediction, Logistic Regression had the maximum accuracy with 62.58% followed by Support Vector Machine with 60.74%, KNN with 57.87% and Naive Bayes with 56.44%.

Model	Score
Logistic Regression	62.58
Support Vector Machines	60.74
KNN	57.87
Naive Bayes	56.44

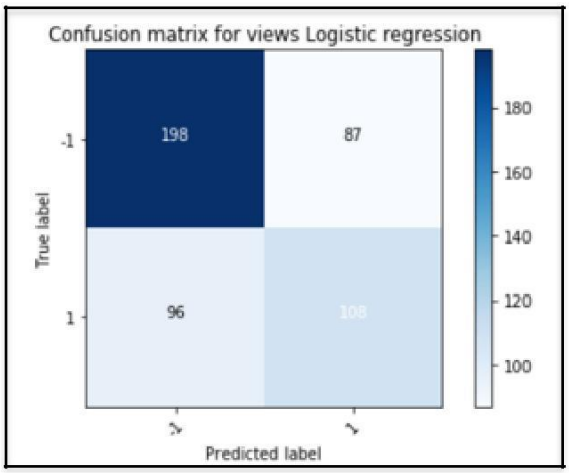
Fig 5. Classification algorithms accuracies for Prediction 1

Precision-Recall is an useful measure of success of prediction in the classification problems. Precision is a measure of result relevancy and Recall is a measure of how many truly relevant results are returned.

The confusion matrix generated shows the precision and recall ratios of the classification algorithms.

Logistic Regression:

The Precision ratio is 0.5538 and Recall ratio is 0.5294.The views lower than the median ie -1 are more in the matrix which are 198.

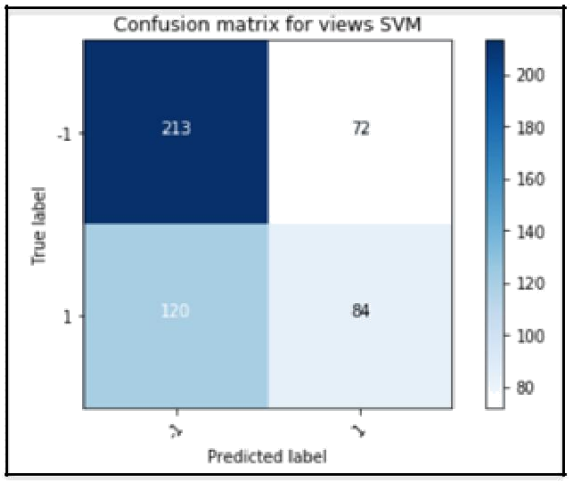



```
Precision : 0.553846153846
Recall : 0.529411764706
F measure : 0.541353383459
```

Fig 6. Confusion Matrix for Logistic Regression

Support Vector Machine :

The Precision ratio for the matrix is 0.5384 and the Recall ratio is 0.4117.

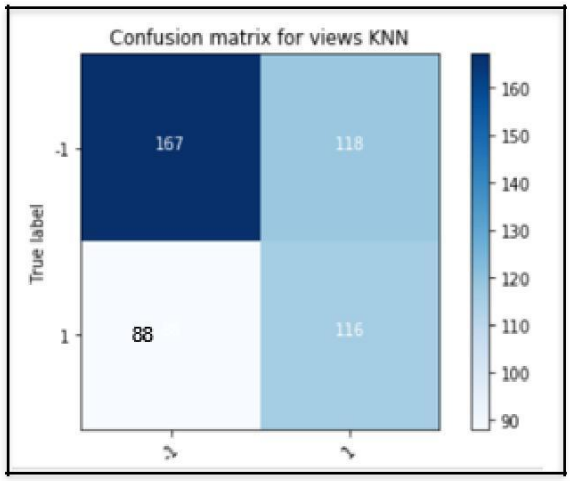


Precision : 0.538461538462
Recall : 0.411764705882
F measure : 0.466666666667

Fig 7. Confusion Matrix for SVM

KNN:

The Precision ratio for the matrix is 0.4957 and the Recall ratio is 0.5686.

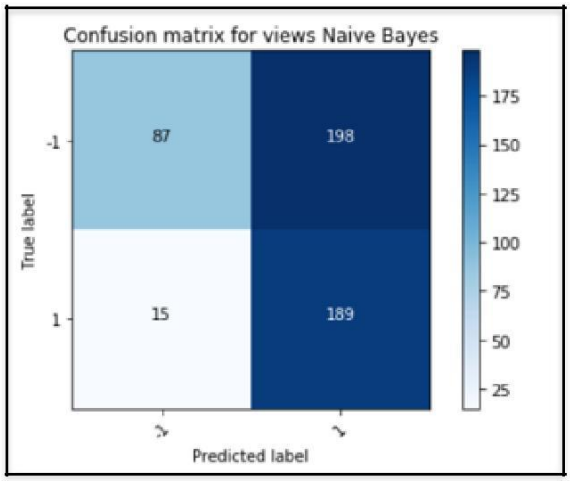


Precision : 0.495726495726
Recall : 0.56862745098
F measure : 0.529680365297

Fig 8. Confusion Matrix for KNN

Naive Bayes:

The Precision ratio for the matrix is 0.4883 and the Recall ratio is 0.9264..



Precision : 0.488372093023
Recall : 0.926470588235
F measure : 0.639593908629

Fig 9. Confusion Matrix for Naive Bayes

For the second prediction where we predicted the sentiments of the talks based on the user review comments, Naive Bayes had the maximum accuracy with 57.06%, Support Vector Machines with 56.85% ,Logistic Regression with 56.44% and KNN with 55.42%.

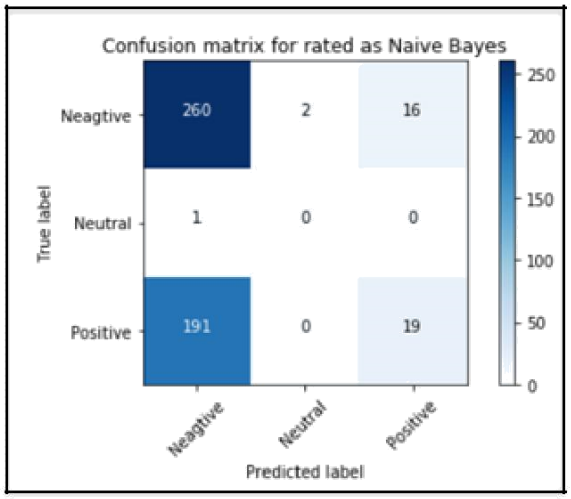
Model	Score
Naive Bayes	57.06
Support Vector Machines	56.85
Logistic Regression	56.44
KNN	55.42

Fig 10. Classification algorithms accuracies for Prediction 2

The confusion matrix generated also shows the precision and recall ratios of the classification algorithms.

Naive Bayes:

The Precision ratio for the matrix is 0.5832 and the Recall ratio is 0.5787.

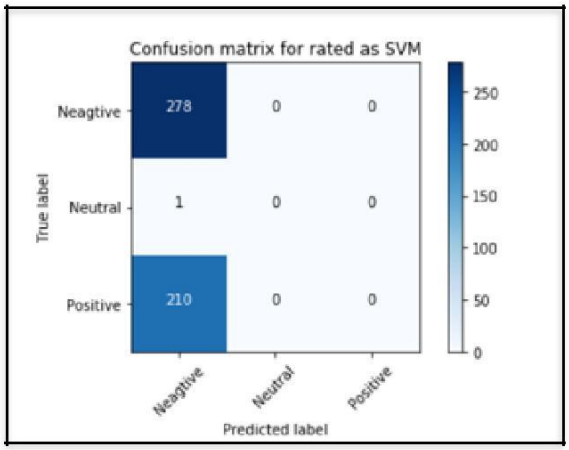



```
Precision : 0.583201307895
Recall : 0.578732106339
F measure : 0.486530957103
```

Fig 11. Confusion Matrix for Naive Bayes

Support Vector Machine:

The Precision ratio for the matrix is 0.3232 and the Recall ratio is 0.5685.

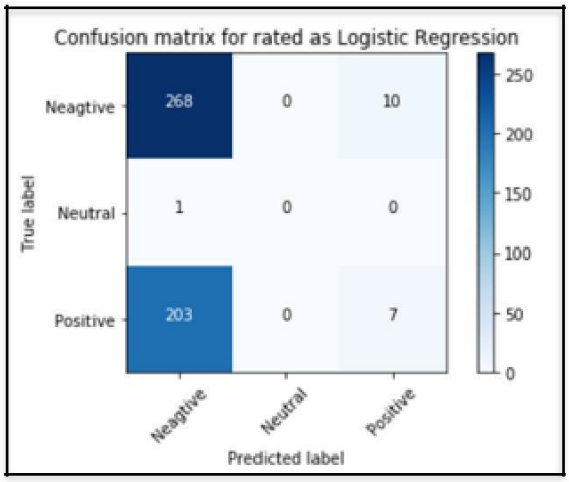



```
Precision : 0.323200388088
Recall : 0.568507157464
F measure : 0.412112098501
```

Fig 12. Confusion Matrix for SVM

Logistic Regression:

The Precision ratio for the matrix is 0.4996 and the Recall ratio is 0.5623.

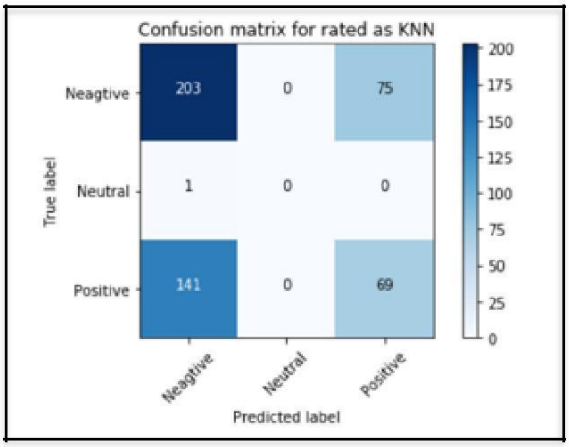



```
Precision : 0.499627905649
Recall : 0.562372188139
F measure : 0.432778885856
```

Fig 13. Confusion Matrix for Logistic Regression

KNN:

The Precision ratio for the matrix is 0.5402 and the Recall ratio is 0.5562.




```
Precision : 0.54029000326
Recall : 0.556237218814
F measure : 0.537899685287
```

Fig 14. Confusion Matrix for KNN

VII. CONCLUSION

We know that if the training set is small with high bias and low variance which is in the case of Naive Bayes, it performs better compared to the classifiers with low bias and high variance like KNN and logistic regression as the latter tends to overfit the model. As the training set grows, the low bias and high variance classifiers become powerful and starts performing better.Based on the results of the classification the best model to predict the number of views is Logistic Regression and the best model to predict the sentiments of the user comments is Naive Bayes.The results for both the predictions are not very high as there always remains several external factors beyond our control that affects the accuracy of the predictions.

REFERENCES

[1] https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ **[2]** http://dataaspirant.com/2016/09/24/classification-clustering-alogrithms/ **[3]** http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[4] www.kaggle.com **[5]** https://www.tableau.com/beginners-data- visualization