Identifying controversial topics in Wikipedia

CSE-506 Data Mining - Course Project

Kshitiz Jain, 2016051

Raj Kamal, 2016076

Dhruv Bhagat, 2016146

Rahul Garg, 2016072

Indraprastha Institute of Information Technology - Delhi

Monsoon 2018

Project Description

Wikipedia articles are one of the widely used encyclopedia. These articles are the combined efforts of many people, in which some are authorized, and rest are anonymous editors. Giving rights to anonymous people gives Wikipedia power to get more detailed and broader overview of the article. But, this also gives rise to the problem of different point of views between editors about the article.

Giving reader the benefit of getting broader coverage about the topic because of plurality of perspectives (Large number of editors), Wikipedia sometimes provides questionable content due to increased controversies between editors.

This creates a need of prior information for reader if the particular Wikipedia article is controversial or not. We have read a number of papers presenting solutions to the same and we proposed a solution in which we take some information about the article's page and its discussion page and classify whether the article is controversial or not.

Data Collection

For the classifiers to work we need to obtain ground truth for training. We have manually obtained the labelled data-set for the classifiers. For controversial topics we took top 500 articles which are claimed by Wikipedia to be most controversial and for the non-controversial topics we took some assumptions like the most worthless articles.

Feature Selection and Pre Processing

In accordance to out literature review we found out that the controversies can be seen in the revision history of discussion pages of article rather than the article itself, therefore we have also taken data of the article's discussion page. We have relied only on statistical data and used the following 7 features as the criteria to classify controversy :

1. revisions of the discussion page
2. minor edits of the discussion page
3. unique editors of the discussion page
4. revisions of the article

5. unique editors of the article
6. revisions of the discussion page by anonymous editors
7. revisions of the article by anonymous editors

Since the labelled data is not available online, we created our own labelled data. For extracting features and creating our dataset we used Wikipedia APIs and MediaWiki-APIs . Since we only get names of top 500 controversial topics we need the url of the respected article to get information about the same. Wikipedia API helps in getting URLs of both the main and discussion pages of the article which are then transferred to the MediaWiki API. MediaWiki API gives a list of all the revisions of the particlar URL, from which we have retrieved the required information.
Following assumptions were taken while obtaining the dataset :

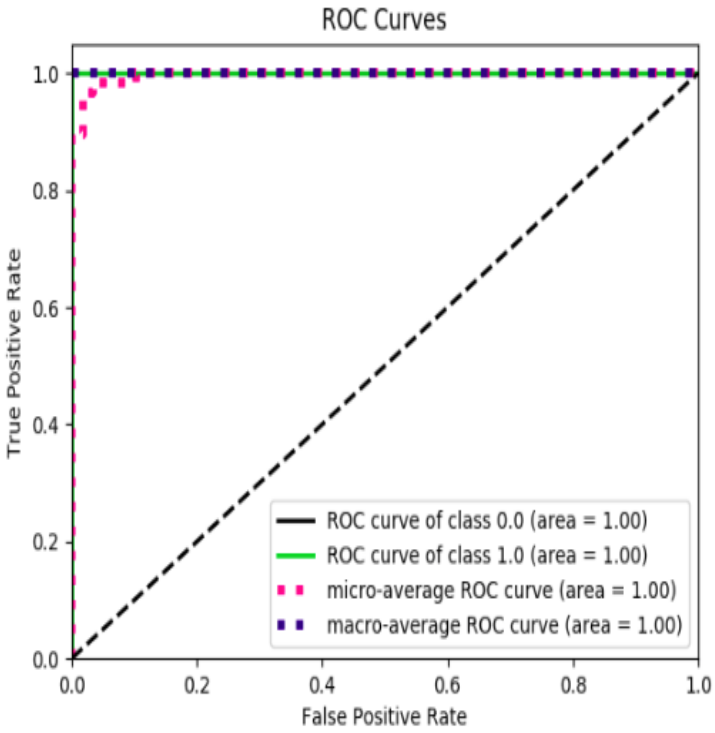
1. If the name of the editor is in numbers i.e. IP of the machine, we have considered it as anonymous edit.
2. Since the names of articles which Wikipedia provides as controversial has discussion page itself as controversial, we have considered it's main page as controversial.
3. Some pages have no anonymous edits, instead the edits with name of "anonymous" or "Anonymous", so the edits corresponding to these names are considered to be anonymous edits.

Classification and results

We have used various classification techniques and obtained results for the same.

0.0.1 Logistic Regression

In logistic regression we find probability of a data point to belonging to a class.The class with the highest probability is chosen as the the class of the data.
Confusion matrix of Logistic Regression
[[5 0]
[2 51]]
Logistic Regression Training Accuracy Score : **0.9770114942528736**
Logistic Regression Testing Accuracy Score : **0.9655172413793104**
Cross Validation Accuracy: **0.95 (+/- 0.10)**

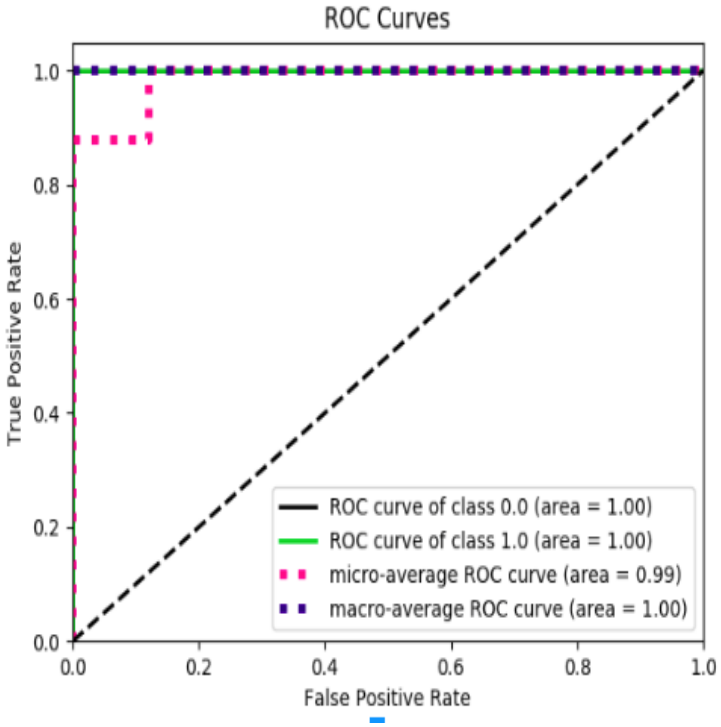


0.0.2 SVM

In SVM we plot the data in n dimensional hyperspace and find the plane that separate the data with maximum margin.
Confusion matrix of SVM
[[7 0]
[0 51]]

SVM Training Accuracy Score : **0.9827586206896551**
SVM Testing Accuracy Score : **1.0**
Cross Validation Accuracy: **0.95 (+/- 0.14)**

The SVM model gives the best accuracy as we are using rbf kernel and using all the data points as features and in our data the points near each other are of the same class.



0.0.3 Neural Network

Neural networks use forward propagation and backward propagation to train the data.
The weights(W) are initially randomised and b are initially zero.Then it is used to find z=W*a[0] +b on which activation function is used.to find a[1] which is used as input for next layer.Similarly it is run for the further layers.After this backward propagation is run where we change our weights to

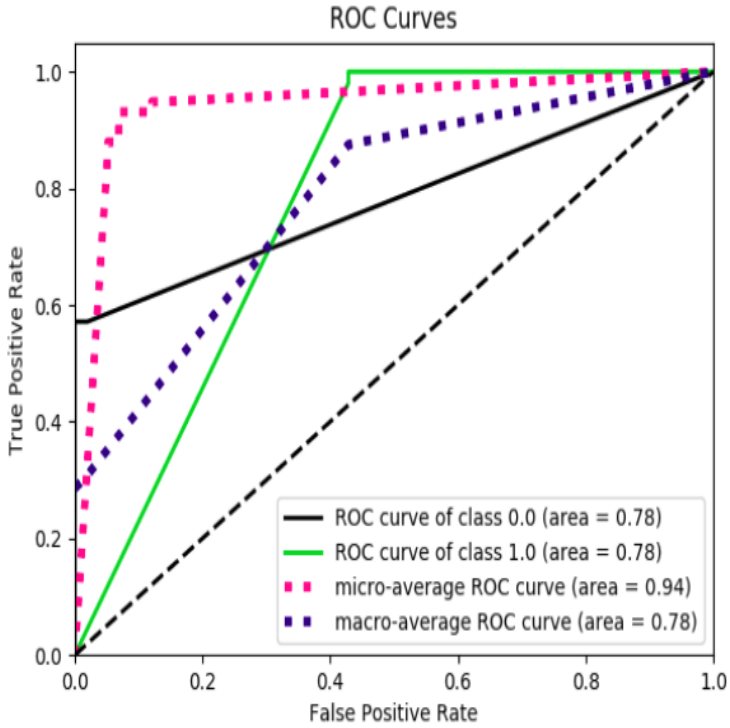
$$w = w - Delta(lossFunction)/Delta(w)$$

and

$$b = b - Delta(lossFunction)/Delta(b)$$

.
We keep on running the above algorithm till convergence reached.
A Neural network of n layers contain n-1 hidden layers.
1 Hidden Layer
Confusion matrix of 1 Hidden Layer Neural Network
[[3 0]
[4 51]]

1 Hidden Layer Neural Network Training Accuracy Score : **0.9080459770114943**
1 Hidden Layer Neural Network Testing Accuracy Score : **0.9310344827586207**
Cross Validation Accuracy: **0.90 (+/- 0.09)**
We can see the accuracy of the model is less than the accuracy of SVM and Logistic Regression this is because with such less data neural network does not perform well.The neural network is under fitting.



3 Hidden Layer
Confusion matrix of 3 Hidden Layer Neural Network
[[2 0]
[5 51]]

3 Hidden Layer Neural Network Training Accuracy Score : **0.9022988505747126**
3 Hidden Layer Neural Network Testing Accuracy Score : **0.9137931034482759**
Cross Validation Accuracy: **0.90 (+/- 0.09)**

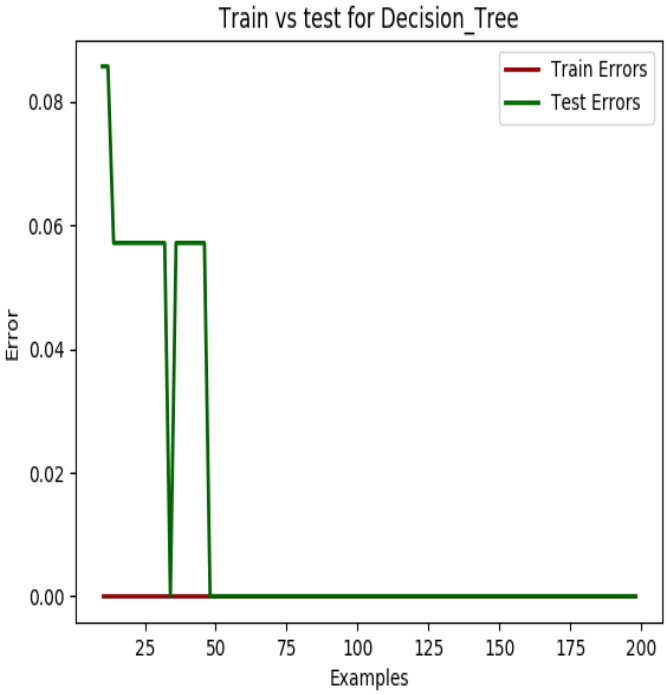
We can see the accuracy of the model is less than the accuracy of SVM and Logistic Regression this is because with such less data neural network does not perform well.The neural network is under fitting.Though the accuracy is better than for

neural network with 2 layers because with more hidden layers we are able to get better features for our data.

0.0.4 Decision Tree

Decision tree function has been used for classification. Decision Tree is tree based classifier each internal node is basically for split testing and each branch represents the outcome of the test (leaf nodes represent the final decisions) It learns by recursive partitioning by making intelligent decisions each time and plotting according until a final classification is achieved.
Confusion Matrix of Decision Tree:
[[6 0]
[0 64]]

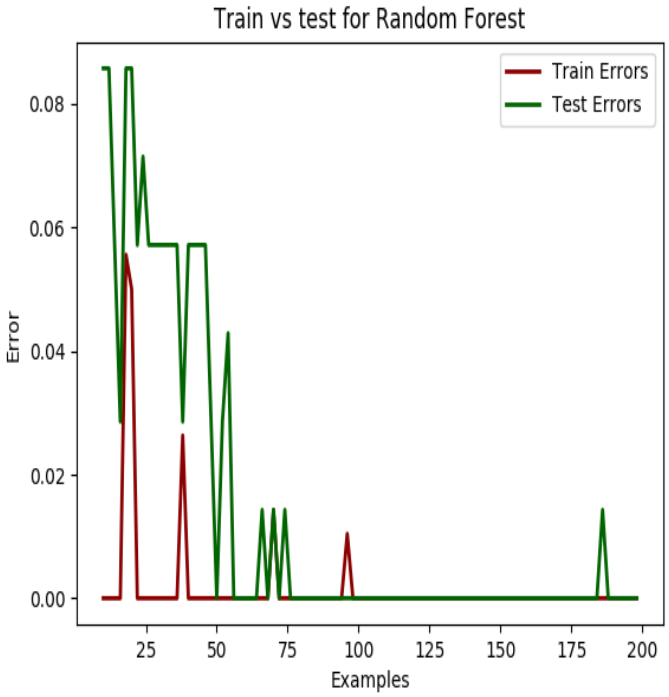
Decision Tree Training Accuracy Score : **1.0**
Decision Tree Testing Accuracy Score : **1.0**
Cross Validation Accuracy: **1.00 (+/- 0.00)**



0.0.5 Random Forest

It basically is an ensemble learning method and basically uses decision trees.It builds an ensemble of decision trees trained and uses merging to build more reliable classifier.Since it adds more randomness,hence its name.
Confusion Matrix of Random Forest:
[[6 0]
[0 64]]

Random Forest Training Accuracy Score : **1.0**
Random Forest Testing Accuracy Score : **1.0**
Cross Validation Accuracy: **0.98 (+/- 0.08)**

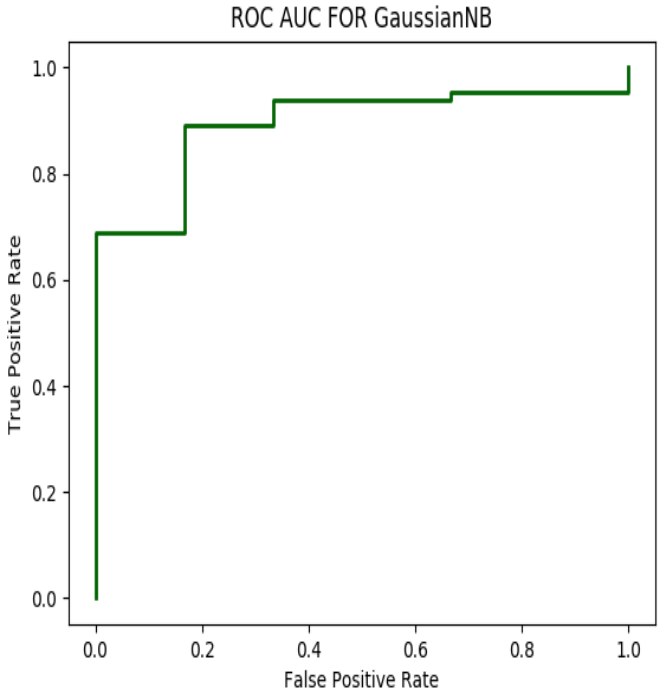


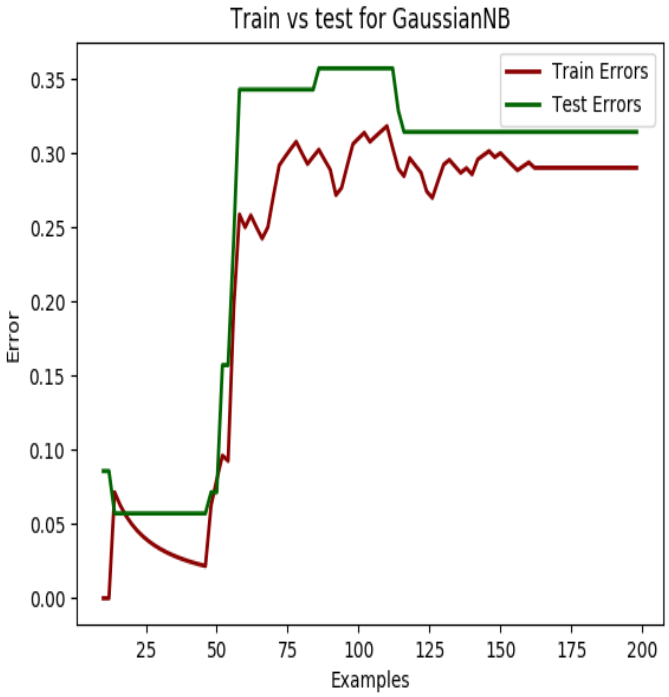
0.0.6 Gaussian Naive Bayes

Conditional probabilities is the basis of the Bayes' theorem which help us to calculate probability whether something will happen or not, based on something already happened. Naive Bayes Classification is based on Bayes' theorem. All features used in the classification are considered independent of each other. Gaussian Naive Bayes is a distinct version of Naive Bayes where the values of the features

are continuous. Normal distribution is considered for conditional probability calculation.
Confusion Matrix of Gaussian Naive Bayes :
[[6 0]
[22 42]]

Gaussian Naive Bayes Training Accuracy Score : **0.7098765432098766)**
Gaussian Naive Bayes Testing Accuracy Score : **0.6857142857142857)**
Cross Validation Accuracy: **0.71 (+/- 0.23)**

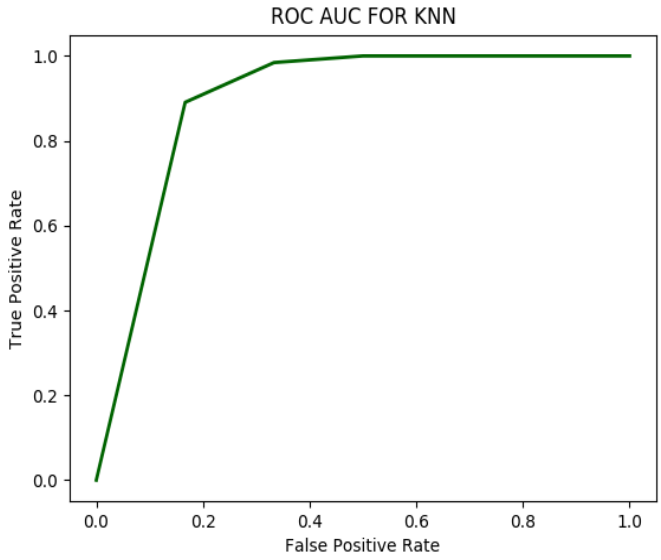


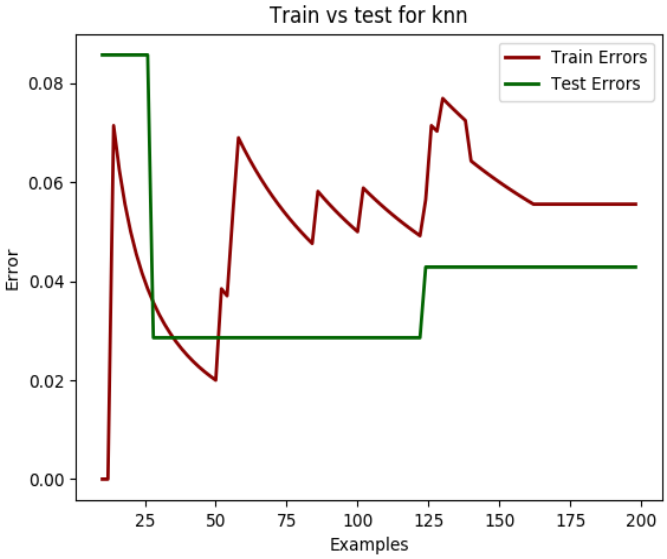


0.0.7 K-Neighbours

To classify data-point a poll/vote is conducted from its neighbouring sets of data-points and assigned to one to which it's most relevant or closest among its k-neighbours.
Confusion Matrix of K-Neighbours:
[[4 2]
[1 63]]

K-Neighbours Training Accuracy Score : **0.9444444444444444)**
K-Neighbours Testing Accuracy Score : **0.9571428571428572)**
Cross Validation Accuracy: **0.91 (+/- 0.15**


ROC AUC FOR KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Train vs test for knn
Train Errors
Test Errors
Error
Examples
0.00
0.02
0.04
0.06
0.08
25
50
75
100
125
150
175
200

Results and Inferences

We observed that decision tree and random forest gave best accuracy this is because decision trees use split testing at every branch, until a leaf node is reached which represents our final decision. Random forest is simply a collection of decision trees whose results are aggregated into final result. The next best classifier for our data is SVM because our data is linearly separable in higher dimensions thus SVM was able to generate a large margin for our data and predict the results. Logistic Regression also works greatly as probability that a wikipedia page is controversial or not comes out to be approximately one for a given data point. K Nearest Neighbour assign the data point to the class to which it's nearest neighbour belongs to.The Neural Network forward propagates and backward propagates in training of data and uses the weights to predict the outcome for our test data. The Naive Bayes Algorithm gives the worst accuracy as it considers all features used as independent which is not true for our data.

Problems faced

Following are some challenges faced during the progress of this project.

- Different authors of this research field were approached so as to obtain a reliable dataset for being used in this project. However, due to lack of response from their side, we had to create our own dataset.
- Since no labelled dataset is available online, we had to create our own labelled dataset for training.
    1. We tried using dump data of Wikipedia containing revisions of all the Wikipedia articles, but the size of the dataset was about 20-30 TB of data which cannot be handled by our laptops or lab PCs.
    2. While using MediaWiki and Wikipedia APIs, each query to the API took around 10 minutes to give the revision list of a page. When dealing with pages with controversies, it took approximately 10 additional minutes to iterate the additional huge lists associated. Hence, creating the whole dataset took around 2-3 days on lab PCs.
- Some assumptions were taken while creating dataset as mentioned in the feature selection section.

References

Following are some libraries and resources we used during the progress of this project.
Libraries:

- Sklearn : Model Selection, Classifier, Confusion Matrix, Calculating Accuracy Score,
- Numpy : Importing Data, Display plots
- Matplotlib : Plotting Train/Test and Accuracy for given Classifier
- Pandas : Reading CSV Data set
- Scikitplot :Plotting ROC Curve
- Urllib2 : Reading URL Pages for data extraction
- Re : Reading the revision history of Wiki-Page in set of 500
- Socket :Validation of String as IPv4 or IPv6
- Itertools: Simultaneous iterating across files
- Wikipedia: Finding URL for a given Wikipedia Article and its Discussion.

Resources

1. Controversial Pages : Wikipedia
2. Non-Controversial Pages : Worthless wikipedia pages
3. Features Selection : Proceedings of the 2014 International Conference on Social Computing.